Computational identification of genes: *ab initio* and comparative approaches

Genís Parra Farré

PhD thesis

Barcelona, December 2004

The Figure in the cover shows a representation of `geneid` predictions (Figure 1, Parra et al. (2000)).

UNIVERSITAT POMPEU FABRA
Departament de Ciències Experimentals i de la Salut

Computational identification of genes: *ab initio* and comparative approaches

Genís Parra Farré

Memòria presentada per optar al grau de Doctor
en Biologia per la Universitat Pompeu Fabra.

Aquesta Tesi Doctoral ha estat realitzada sota la direcció del
Dr. Roderic Guigó Serra al Departament de Ciències Experimentals
i de la Salut de la Universitat Pompeu Fabra

Roderic Guigó Serra **Genís Parra Farré**

Barcelona, December 2004

The research in this thesis has been carried out at the Genome Bioinformatics Lab (GBL) within the Grup de Recerca en Informàtica Biomèdica (GRIB) at the Parc de Recerca Biomèdica de Barcelona (PRBB), a consortium of the Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra (UPF) and Centre de Regulació Genòmica (CRG).

The research carried out in this thesis has been supported by grants from Ministerio de Ciencia y Tecnología to R. Guigó.

To my parents
To my brother and sisters

Motivation

It is clear that we are living a really important period in the development and the knowledge of life sciences. Fifty years after the description of the structure of the double helix, we have moved from the analysis of a single gene to the systematic mass sequencing of entire genomes. At the time of writing this dissertation, whole genome sequencing projects for hundreds of organisms (bacteria, archea and eukaryota, as well as many viruses and organelles) are either complete or underway. All the information we are gathering today will probably modify the way we will understand life, science and medicine. But, before the best use can be made of this data, the identification and the precise location of the functional regions of the genomic sequences must be determined. The most important things to realize about the "book of life", is that we know almost nothing about the language in which it is written, and that raw genomic sequences are mainly useless for the scientific community. The challenge ahead is to extract relevant information encoded within the billions of nucleotides stored in our databases.

In a very simplistic description, the first step in the functional annotation of a genome would be to find the collection of genes encoded in the nucleic acid sequences. The next step would be to assign a function to each protein, where the three dimensional structure of the proteins will play a key role. Then, using microarrays technology, it will be feasible to obtain the spatial and temporal expression pattern of each gene at any developmental stage or specific condition. Finally, the last step would be to establish the network of interactions and regulations among all the proteins of a complete genome.

This thesis focuses on the first step of any genome analysis: to find where genes are. The motivation of this thesis, thus, is to give a little insight in how genes are encoded and recognized by the cell machinery and to use this information to find genes in unannotated genomic sequences. The complexity of gene prediction differs substantially in prokaryotic and eukaryotic genomes. While prokaryotic genes are encoded in single continuous open reading frames, usually adjacent to each other, eukaryotic genes are separated by long stretches of intergenic regions, and their coding sequences can be interrupted by large non coding sequences. One of the objectives of this thesis is the development of tools to identify eukaryotic genes through the modeling and recognition of their intrinsic signals and properties.

This thesis addresses another significant open problem of this field: how the sequence of related genomes can contribute to the identification of genes. The value of comparative genomics is illustrated by the sequencing of the mouse genome for the purpose of annotating the human genome. The availability of closely related genomes makes it possible to carry out genome-wide comparisons and analysis of syntenic regions. Recently, compar-

ative gene predictions programs exploit this data under the assumption that conserved regions between related species correspond to functional regions (coding genes among them). Thus, the second part of this thesis describes a gene prediction program that combines *ab initio* gene prediction with comparative information between two genomes to improve the accuracy of the predictions.

Nowadays computational analysis is a major, integral part of genomics. It would not be an exaggeration to claim that genomics analysis can only be made with computational tools. Only by using computational methods and statistical models we can try to find out how genes are encoded and try to accurately predict their location in complete genomic sequences.

Thus, the work described in this dissertation is essentially interdisciplinary; this means that while the basic subject of matter is biological and the obtained results are of biological interest, techniques from other fields have been extensively used. Statistical approaches have been used to create models of genomic features to be able to recognize sequence motifs and reproduce the underneath biological process, while computational programming has been applied to include these models into efficient bioinformatic tools.

Genís Parra

Barcelona, December 2004

Acknowledgments

It is not just to follow convention that I first acknowledge my PhD advisor, Roderic Guigó. Quite simply, if not for him, my academic career would have finished with my bachelor degree. It was he who saw past my sub-optimal scores to someone able to work on research. I am indebted to him for letting me start what I hope will be a long career in research.

Other people thath I would like to thanl are: Pankaj Agarwall, for the stage in the GlaxoSmithKline, the Dyctiostelium annotation group: Karol Szafranski, Gernot Glökner and Mathias Platzer, the people from the University of Geneva: Manolis Dermitzakis, Alexandre Reymond and Stylianos Antonarakis and Michael Brent and all the people of his lab for the scientific collaboration and the invitation to Saint Louis.

I would also like to mention the people who may not had a direct impact on this thesis, but have influenced it indirectly by molding me into the person I am today. All the people who have given of their time, talents, and expertise to help me on this project have enriched my life. For their special friendships and assistance, I am most grateful:

To Mercè for being there in the darkest years of my PhD (and life). For encouraging me when I was giving everything up. For your trust in friendship. For the years we lived together and for all the amazing things you taught me. For all the incredible journeys we did. For all the love you gave me.

To Sergi for those Sunday afternoons in la filmoteca. For your way of living life and science. For la Passió d'Esparraguera, Eric Sardinas and for your comics. For those nights in New York. Special thanks for your cocktails, for listening to me and for your wise advises.

To Cristina for your enthusiasm and your energy. For your tiramisu, for your pesto and for your profiteroles. For all your tenderness and comprehension. For Patrizio, Gurdieff and Dilan Dog. For going with me to fill my bottle of water every day. For the Ravenna mosaics. For bringing happiness and joy in our every day life.

To Robert for your thesis template, for your whiskey and Risk sessions. For the parties on your roof. For being our volleyball captain. For your pictures of California (that decided my future). For your wise statements.

To Pep for sharing a fraction of all your infinite wisdom with me. I learned (or at least I tried) from you to try to do the things the best one can.

To Enrique for programing `geneid`. I learned a lot while working with you. For your patience with my `geneid` problems. For all the course we teached and the moments we shared.

To Fabien and Isabel for those roller hockey nights. For your penguin. For your strength and courage. For your friendship.

To Xavi for being our mentor in the early days. I will say nothing about your home directory. For being a destroyer. Quin payo !!

To Bet for massaging my breast. For gifts you give. For your sympathy and friendship. For your complicity.

To Rut for the swimming mornings, for the theater, for your smile. Remember me when you become a famous actress.

To Noura for being just like you are. For your voice and for that night in the karaoke.

To Ramon for all those amazing gadgets you have. For your outdoor activities. For your true friendship, for your music, for your stories and for el Pilar.

To Peppolino for your hugs, for running with me, for the musica pertarda, and for the Arena sessions.

To Citlali for all the pushes and shoves playing hockey. For Valencia and Blanes. For how I feel being with you. For your guacamole.

To Moisès for installing Slackware on my first hard disk and for your long discussions, for your spontaneity and freshness.

To Charles for incite, for your comics and your sense of humor.

To Oscar for all the conversations we had while other people were dancing. For the amazing physical properties of liquids falling inside glasses.

To my students. Specially to my first group: Jimena, Encarni, Bet and Jordi, who show me how difficult is to be a teacher. Just kidding !! You were the best group I ever had. To Gus.

To Jan-Jaap for all the effort you did in the correction of this thesis, it was really a lot of work and I really appreciate it !!!

To Queviures Murgadella for all the food I shared with my friends in the lab.

To Pedro for the organization of the Gulbenkian courses and your kind invitations. I really enjoyed Lisbon and the Gulbenkian courses, and I learned a lot.

To the beach volleyball team that I really enjoyed playing with.

To all the people in the lab that helped me or just supported me, Alfons, Juan Antonio, Miguel, Jorge, Cherraiz, Nicolas, Mar, Hugo, Jordi, Montse, Montse, Francisco and Adrian.

To the ecuador gang: Xavi, M Jose, Manuela, Lourdes, Jordi, Francesc, Laia, Ethel and Victor. For the amazing moments around the world.

To the PhD. courses mates : Susana, Miki, Aida, Clara, Anna and David.

To Xavier + for the ride on your boat.

And not to forget the gems of the crown: my friends Albert, Mingo and Joan for being always there and for suffering my incompressible biological talks and my stress. For the mountains we climbed, for the tracks we skied and for the roads we cycled. Thanks for listening to me.

Without all of you this thesis would not have been possible. Thanks again.

Contents

List of Tables

List of Figures

Introduction

1.1 Biological background

In the initial part of this chapter we try to define what a gene is. This is not an easy task and this section owes its overall structure to the recent review done by Snyder and Gerstein (2003). The second part of the section describes some of the molecular processes that are involved in the pathway from the nucleic acids to proteins.

1.1.1 What is a gene?

Historically, the term gene (from the Greek word *genos*, which means birth, race or offspring), was coined by the Danish botanist Wilhelm Johannsen in the early 1900s as an abstract concept to explain the hereditary basis of traits. He also made the distinction between the morphological appearance of an individual (phenotype) and its genetic traits (genotype).

Earlier, William Bateson, a geneticist supporter of Mendel's ideas, had used the word genetics in a letter; he felt the need for a new term to describe the study of heredity and inherited variations. But the term did not start spreading until Wilhelm Johannsen suggested the Mendelian factors of inheritance to be called genes.

Phenotypic traits were associated with hereditary factors even though the physical basis of those factors were not known. Early genetic studies by Thomas Hunt Morgan and others associated heritable traits with specific chromosomal regions. Using fruit flies as a model organism, Thomas Hunt Morgan and his group at Columbia University showed that genes, located in the chromosomes, are the units of heredity. In 1930s, George Beadle, based on mutagenesis experiments, introduced the concept of "one gene, one enzyme", which later became "one gene, one protein".

With the development of recombinant techniques and gene cloning, in the 60s, it became possible to combine the assignment of a gene to a specific segment of a chromosome and the synthesis of a gene product. Although it was originally presumed that the final product was a protein, the discovery that ribonucleic acids can have structural, catalytic, and even regulatory properties made it evident that the end product could be a nucleic acid. Thus, we can define a gene in molecular terms as "a complete chromosomal segment responsible for making a functional product". This definition has several logical compo-

nents: the expression of a gene product, the requirement that it must be functional, and the inclusion of both coding and regulatory regions.

Nowadays, however, in the so-called "genomic era", geneticists have realized that the complexity of the genomic information is even beyond the standard definition of gene. There are additional "peculiarities" in gene identification that do not fit in previous definitions (i.e. overlapping transcripts, alternative splicing, fusion genes or pseudogenes).

There are now examples of overlapping transcriptional units (for example where exons of one gene are encoded within the intron of another) and even overlapping protein coding genes (Coelho et al., 2002; Tycowski et al., 1996). In all cases of gene overlap, each gene has a unique functional sequence that is different from the others.

The production of several isoforms from the same transcriptional unit by various types of alternative splicing seems to be a very common event. A single primary transcript can have several regions that can generate alternative splicing. Thus, the resulting combinatorial effects of selecting different splice sites can be very pronounced, and genes that code for tens to hundreds of different isoforms could be common (Graveley, 2001). In the human genome, at least half of all genes have alternative spliced isoforms and this is likely to be an underestimation because not all transcript variants have been identified (Modrek and Lee, 2002).

There is also evidence that, in some cases, two adjacent genes are transcribed together. Genes in prokaryotic organisms are organized into operons that generate long transcripts encoding for many proteins that are transcribed together and translated sequentially. The phenomenon in eukaryotic genomes that we are describing involves the synthesis of one protein from two fused genes. Thus, an authentic gene fusion event would possess a particular mechanism to override the nonsense codon that would be used to stop translation of the protein from the first gene. Employing the efficient splicing system in eukaryotes, a few observed cases use alternative splicing to skip the exons containing the stop codon. Therefore, there are described cases where genes from two adjacent loci are transcribed together, probably as a result of a weak terminating signal, and after splicing, a fusion mature transcript is built skipping the stop signal and generating a new chimeric protein with exons from both pre-existing and independent genes (Thomson et al., 2000; Poulin et al., 2003).

The definition of a gene is also linked with the definition of a pseudogene. Pseudogenes are similar in sequence to normal genes, but they usually contain obvious disablements such as frame shifts or stop codons in the middle of coding domains. This prevents them from producing a functional product or having a detectable effect on the organism's phenotype. The boundary between "living" and "dead" genes is often not so sharp. Pseudogenes can be transcribed; truncated proteins could still have some functionality, and regions with stop codons could be alternatively spliced. Conversely, there are some pseudogenes that have entire coding regions without obvious disablements but do not appear to be expressed. Presumably they lack the regulatory elements required for the transcription. In these cases, it is difficult or almost impossible to discard marginal transcription in some isolate tissue at some developmental stage.

As we have seen, the term gene has a broad and often diffuse definition. This plasticity in the way that genes are specified could be convenient for the cell in order to generate a huge amount of different combinations of final mRNAs and subsequently a huge amount of protein diversity. For the rest of this thesis, the term gene will be used to refer to protein

coding regions, and to simplify the problem, overlapped transcripts, fusion transcripts and alternative splicing isoforms will not be taken into account.

1.1.2 Molecular basis of genomic information

Deoxyribonucleic acid (DNA) is a double-stranded molecule that is twisted into a spiral staircase like helix. Each strand is composed of a sugar-phosphate backbone and numerous base chemicals attached in pairs. The four bases that make up the steps in the spiraling staircase are adenine (A), thymine (T) (uracil (U) when we are referring to the ribonucleic acid, RNA), cytosine (C) and guanine (G). These steps act as the "letters" in the genetic alphabet, combining into complex sequences to form the words, sentences and paragraphs that code for the instructions to guide the formation and the differentiation of the cell. Maybe even more appropriately, the A, T, C and G in the genetic code of the DNA molecule can be compared to the "0" and "1" in the binary code of computer software. Like software to a computer, the DNA code is a genetic language that communicates information to the organic cell. It was not until the early 90s and the Human Genome Project that the scientific community deeply began to explore the nature and complexity of the digital code inherent in DNA and bioinformatics rose as the only way to provide tools to manage this type and amount of information.

But, returning to the biological problem, how do we move from DNA to proteins? This process, known as the central dogma of biology, involves three main steps: transcription, RNA processing (including capping, splicing and polyadenylation) and translation. The basic schema of the central dogma is shown in Figure 1.1.

Transcription

As the initial step in gene expression, transcription is the central point of many regulatory mechanisms. Eukaryotic genes contain highly structured regulatory sequences that direct complex patterns of expression in many cell types during different stages of development. The transcription of a gene is modulated by the interactions between specific proteins that bind regulatory elements in the genomic sequence. These proteins function as transcription factors needed for the RNA polymerase to initiate transcription. The control region combines several different kinds of regulatory elements, and suggests the principle that when a promoter is regulated in more than one way, each regulatory event depends on the binding of its own protein to a particular sequence. When the optimal combination of transcription factors are bound to their corresponding sequence elements, the continuous sequence of DNA corresponding to a single gene is copied to an RNA sequence by the RNA polymerase II.

The degree of complexity of the transcriptional regulatory regions differs notably among eukaryotes and seems to correlates with structural and behavioral complexity. A typical yeast regulatory region consists of short sequences located immediately upstream of the transcription start site (see Figure 1.2 **a**). Most core promoters contain a TATA element, which serves as a binding site for TBP (TATA-binding protein). In general, promoters are selected for expression by the binding of TBP to the TATA element. The regulation of the TBP binding depends on upstream activating sequences, which are

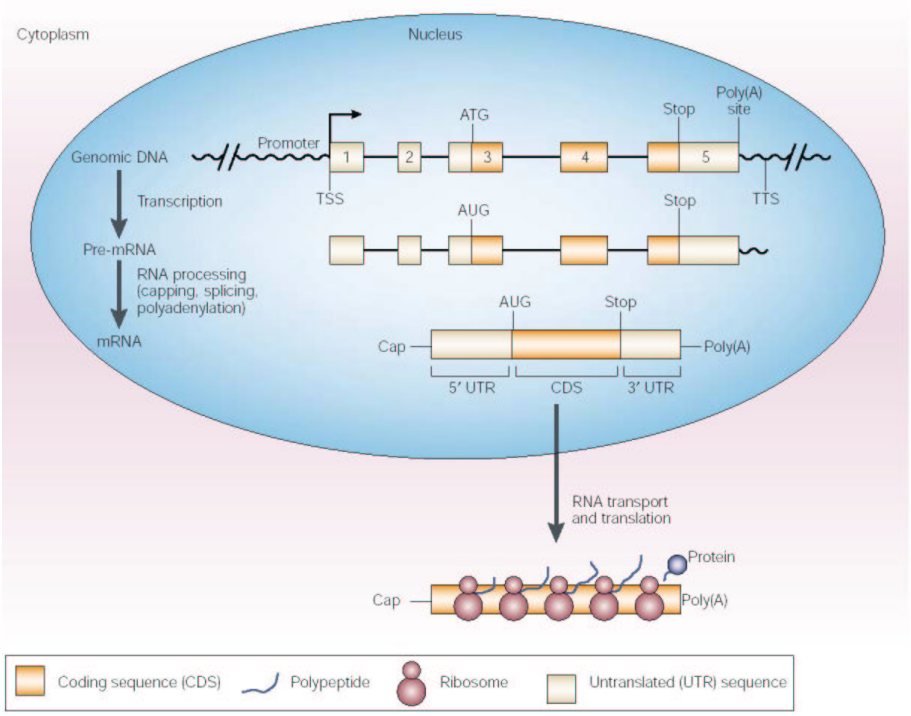


Figure 1.1: Schema of the central dogma of gene expression. In the typical process of eukaryotic expression, a gene is transcribed from DNA to pre-mRNA. mRNA is then produced from pre-mRNA by RNA processing, which includes the capping, splicing and polyadenylation of the transcript. It is then transported from the nucleus to the cytoplasm for the translation. From Zhang (2002).

usually composed of 2 or 3 closely linked binding sites for one or two different sequence-specific transcription factors. A few genes in the yeast genome contain distal regulatory sequences, but the majority contains a single upstream activating sequence located within a few hundred base pairs of the TATA element (Levine and Tjian, 2003).

A typical metazoan gene is likely to contain several enhancers that can be located in 5′ or 3′ regulatory regions, as well as within introns (see Figure 1.2 **b**). Each enhancer is responsible for a subset of the total gene expression pattern; they usually mediate expression within a specific tissue or cell type. A typical enhancer is something like 500 bp in length and contains in the order of ten binding sites for at least three different activators and one repressor (Levine and Tjian, 2003). The core promoter is compact and extends few hundred bases upstream of the transcription start site. There are at least three different sequence elements that can recruit the TBP complex: the TATA element, the initiator element and the downstream promoter element (DPE). Core promoters that lack a TATA sequence usually containing a compensatory DPE element, in order to ensure recognition by the RNA polymerase II transcription complex.

Many genes contain binding sites for proximal regulatory factors located just 5′ of the core promoter. These factors do not always function as classical activators or repressors; instead, many of them work as recruiting elements for distal enhancers to the core promoter. Finally insulators prevent enhancers associated with one gene from inappropriately regulating neighboring genes. These regulatory genomic sequences: enhancers, silencers and insulators, are scattered over distances of roughly 10 Kbp in fruit flies and 100 Kbp in mammals (Levine and Tjian, 2003).

A dominant characteristic of promoter sequences in the human genome is the abundance of CpG dinucleotides. Methylation plays a key role in the regulation of gene activity. Within regulatory sequences, CpGs remain unmethylated, whereas up to 80% of CpGs in other regions are methylated on a cytosine. Methylated cytosines are mutated to adenosines at a high rate, resulting in a 20% reduction of CpG frequency in sequences without a regulatory function as compared with the statistically predicted CpG concentration (Fazzari and Greally, 2004). CpG islands have been identified at the promoter sites of approximately half of the gene in the human genome, most of which are considered to be “house keeping” genes according with their ubiquitous expression pattern.

Transcription termination by RNA polymerase II seems to be only loosely specified. In some transcription units termination occurs beyond 1000 bp downstream of the site corresponding to the mature 3′ end of the primary transcript (which is generated by cleavage at a specific sequence). Instead of using a specific terminator sequence, the enzyme stops RNA synthesis within multiple sites located in rather long “terminator regions” (Lewis, 1997). The nature of individual termination sites is not known.

RNA processing

There are three main RNA modifications: the capping reaction, splicing and the maturation of the 3′ end by cleavage and polyadenylation. These three processes occur while the RNA is being synthesized. There is evidence that regulatory interactions among these processes and transcription (through the C-terminal domain of the RNA polymerase II) are crucial to obtain the final mature RNA. Recent studies have shown that the “mRNA

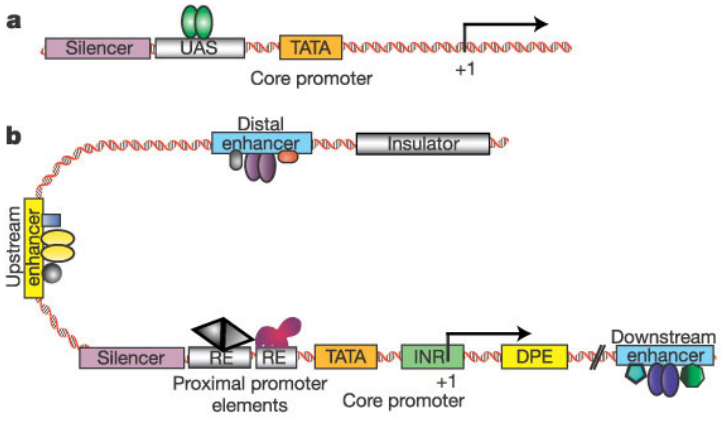


Figure 1.2: Comparison of a simple eukaryotic promoter and a extensively diversified high eukaryotic regulatory modules. **a)** Simple eukaryotic transcriptional unit. A simple core promoter (TATA), upstream activator sequence (UAS) and silencer element spaced within 100-200 bp of the TATA box that is typically found in unicellular eukaryotes. **b)** Complex metazoan transciptional control modules. A complex arrangement of multiple clustered enhancer modules interspersed with silencer and insulator elements which can be located 10-50 kb either upstream or downstream of the core promoter containing TATA box initiator sequences (INR), and downstream promoter elements (DPE). Adapted from (Levine and Tjian, 2003).

factory" is a dynamic complex whose composition changes as it moves along the transcribed sequence of genes (Zorio and Bentley, 2004).

Capping

The 5′ end of the RNA (which is the end synthesized first during transcription) is capped by the addition of a methylated G nucleotide. Capping occurs almost immediately, after about 30 nucleotides of RNA have been synthesized, and it involves condensation of the triphosphate group of a molecule of GTP with a diphosphate left at the 5′ end of the initial transcript. The new G residue added to the end of the RNA is in reverse orientation from all the other nucleotides. This 5′ cap will later play an important part in the initiation of protein synthesis and it also seems to protect the growing RNA transcript from degradation (Lewis, 1997).

Polyadenylation

The 3′ ends of mRNAs are generated by cleavage followed by polyadenylation. RNA polymerase transcribes past the site corresponding to the 3′ end, and sequences in the RNA are recognized as targets for an endonucleolytic cut followed by polyadenylation. A common feature of the mature transcripts in higher eukaryotes (not including yeast) is the presence of the sequence AAUAAA in the region from 11-30 nucleotides upstream of the site of poly(A) addition. The sequence is highly conserved and only occasionally is even a single base different. Deletion or mutation of the AAUAAA hexamer prevents generation of the polyadenylated 3′ end. The signal is needed for both cleavage and polyadenylation (Lewis, 1997). Generation of the proper 3′ terminal structure requires an endonuclease (consisting of the cleavage factors CFI and CFII) to cleave the RNA, a poly(A)polymerase (PAP) to synthesize the poly(A) tail, and a specificity component (CPSF) that recognizes the AAUAAA sequence and directs the other activities.

The addition of poly(A) helps to stabilize the mRNA and seems to be related with the efficiency of translation initiation. The average size of the poly(A) tail is from over 70 adenosines in yeast, to over 240 adenosines in mammals for newly transcribed mRNA and pre-mRNA in the nucleus. Cytoplasmic enzymes may also cause the polyA to shorten, and occasionally lengthen before translation. Not all transcripts are polyadenylated; some histone mRNA is poly(A) negative. For transcripts without poly(A) tail, the 3′ end seems to be protected or sequestered by association with other factors.

Splicing

The primary RNA transcript is spliced to remove intron sequences, producing a shorter RNA molecule. Introns are removed from the nuclear RNAs of eukaryotes by a system that recognizes short consensus sequences conserved at exon-intron boundaries and within the intron. The splicing of precursors to mRNAs occurs in two steps, both involving single transesterification reactions. The first step generates a 2′-5′ bond at the branch site upstream of the 3′ splice site and a free 3′ hydroxyl group on the 5′ exon generating a lariat RNA intermediate. The second step involves an attack of the 3′ hydroxyl group on the phosphodiester bond at the 3′ exon and results in the joining of the two exons.

This reaction requires a large splicing apparatus, which takes the form of an array of proteins and ribonucleoproteins that generate a large particulate complex, the spliceosome. There are two distinct types of spliceosome in most organisms. The major class or U2-type (also known as canonical or GT-AG splice sites) is universal in eukaryotes, whereas the minor class or U12-type (also known as AT-AC) may not be present in some organisms. The consensus sequences of U12-type introns are more conserved than those of vertebrate U2-type introns (Sharp and Burge, 1997). Although less conserved, the U2-type involved signals still have clearly recognizable motifs (see Figure 1.3).

The U2-type spliceosome is composed of five small nuclear RNAs (snRNAs) called U1, U2, U4, U5, and U6 and numerous protein factors. Splice site recognition and spliceosomal assembly occur simultaneously according to a complex sequence of steps (see figure 1.4). The first step appears to be the recognition of the donor (5′) splice site at the exon-intron junction: a substantial amount of genetic and biochemical evidence has established that this occurs primarily through base pairing with the U1 snRNA over a stretch of approximately nine nucleotides, including the last three exonic nucleotides and the first six nucleotides of the intron. The second step in spliceosomal assembly involves binding of U2 auxiliary factor (U2AF) and possibly other proteins to the pyrimidine-rich region immediately upstream of the acceptor site, which directs U2 snRNA binding to the branch point sequence approximately 20 to 40 bp upstream of the intron-exon junction. The U2 snRNA sequence 3′ GGUG 5′ has been show to base pair with the branch point signal, consensus 5′ YYRAY 3′, with the unpaired branch point adenosine outstanding of the RNA duplex. Mutations or deletions of the branch site in yeast prevent splicing. In higher eukaryotes, the relaxed constraints in its sequence result in the ability to use related sequences in the vicinity when the authentic branch site is deleted. Subsequently, a particle containing U4, U5, and U6 is added to the spliceosome. The subunit U5 possibly interacts with the acceptor site, leading eventually to the formation of the mature spliceosome.

Several examples of intronic and exonic cis-acting elements, important for correct splice site identification and distinct from the classical splicing signals, have been described recently. These elements can act stimulating (as enhancers) or repressing (as silencers) splicing, and they seem to be especially relevant for the regulation alternative splicing. Exonic splicing enhancers (ESEs) in particular appear to be very prevalent and might be present in most, if not all, human exons, including constitutive ones (Cartegni et al., 2002). The lack of a well-defined consensus sequence for these signals indicates that they might consist of numerous functionally different classes, and that the factors involved may recognize degenerate signal sequences (Cartegni et al., 2002).

Translation

The mRNA sequence is translated into protein sequence, outside the nucleus, by a subcellular structure known as ribosome (a compact ribonucleo-protein consisting of two subunits). The ribosome binds to the mRNA, and scans the sequence synthesizing the amino acid sequence specified by consecutive non-overlapping codons. Codons are defined as triplets of consecutive nucleotides which are recognized by the transfer RNA (tRNA) with the corresponding attached amino acid. Scanning of the mRNA proceeds until the ribosome finds one of the three codons not specifying amino acids (the stop codons: UGA, UAG and UAA). At that point, elongation of the amino acid sequence ends and the final protein product is released.

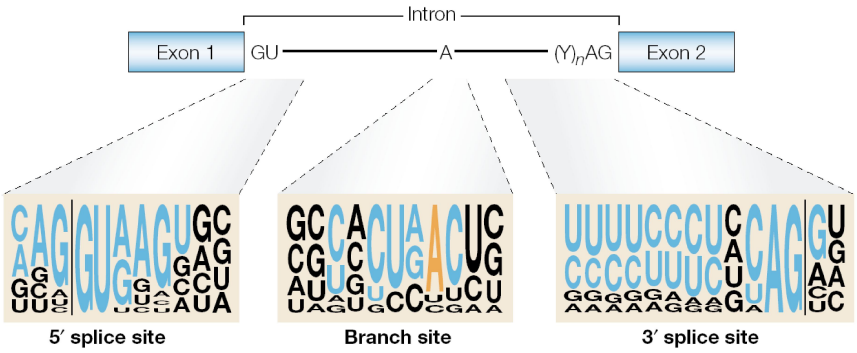


Figure 1.3: Splicing sequence motifs for U2-type spliceosome. The nearly invariant GU and AG dinucleotides at the intron ends, the poly-pyrimidine tract preceding the 3′AG, and the A residue that serves as a branch point are shown. For each sequence motif, the size of nucleotide at a given position is proportional to the frequency of that nucleotide at that position in an alignment if conserved sequences. Nucleotides that are part of the classical consensus motifs are shown in blue, except for the branch point A, which is shown in orange. Adapted from Cartegni et al. (2002).

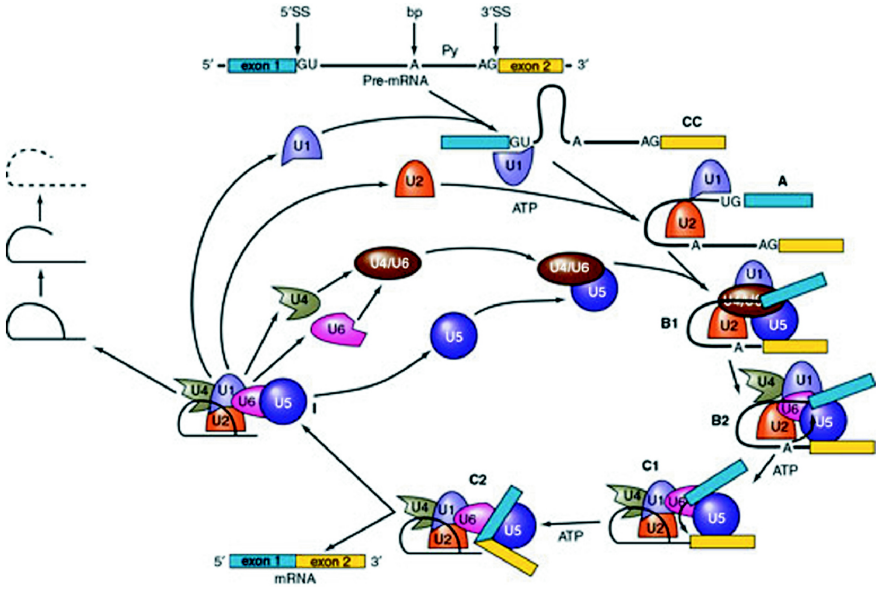


Figure 1.4: The spliceosome cycle. The processing of the pre-mRNA containing two exons and one intron into the ligated exon product and lariat intron is shown, emphasizing the involvement of the small nuclear ribonucleoprotein (snRNP) particles at distinct steps in spliceosome formation and catalysis. Adapted from Burge et al. (1999).

Selection of the start codon, a methionine codon triplet, sets the reading frame that is maintained normally throughout all subsequent steps in the translation process. What makes the start different from the addition of a methionine internally in the polypeptide chain is a special tRNA, initiator transfer RNA (tRNAi), that is used to recognize the translational start codon. When this tRNAi is charged with Met to form met-tRNAi, this compound binds to the P site of ribosomes. In eukaryotes, the small (40S) ribosomal subunit carrying met-tRNAi and other associated proteins recognizes the 5′ capped end of the mRNA. After the initial recognition it migrates through the 5′ untranslated region (UTR) until it encounters the first "suitable" AUG codon which is recognized by base pairing with the anticodon in met-tRNAi. When a 60S ribosomal subunit joins the paused 40S subunit, selection of the start codon is fixed.

Flanking sequences modulate the efficiency with which the AUG codon is recognized as a stop signal during the scanning phase of initiation. In vertebrate mRNAs, initiation sites usually conform to all or part of the so called translation (or Kozak) signal: GCCAC-CaugGCG (Kozak, 1987).

For maximum effectiveness, the upstream GCCACC motif must directly precede the AUG codon. If the motif is further upstream or the sequence is not optimal the effectiveness in the translation is reduced and even other cryptic AUG codons can be used instead the real one (Kozak, 1999). How the consensus sequence is recognized is not yet known. One possibility is that interaction with GCCACC might slow scanning and thus facilitate the recognition of the AUG codon by met-tRNAi.

Although context effects on AUG codon recognition have been studied primarily in mammalian systems, a strong contribution of the motif has also been demonstrated in plants. In *S. cerevisae*, however, the effects of context are minimal (Kozak, 1999).

Proteins are assembled by the sequential addition of amino acids in the direction from the N-terminus to the C-terminus as a ribosome moves along the mRNA (see Figure 1.5). The genetic code consists of 64 triplets of nucleotides. With three exceptions, each codon encodes for one of the 20 amino acids used in the synthesis of proteins . That produces some redundancy in the code: most of the amino acids being encoded by more than one codon (see Table 1.1). The amino acid is attached to the appropriate tRNA by an activating enzyme (one of 20 aminoacyl-tRNA synthetases) specific for that amino acid as well as for the tRNA assigned to it. An aminoacyl-tRNA (a tRNA covalently bound to its amino acid) able to base pair with the next codon on the mRNA arrives at the A site associated with an elongation factor. The preceding amino acid (Met at the start of translation) is covalently linked to the incoming amino acid with a peptide bond. Then, the ribosome moves one codon downstream alowing the next codon to be binded by the corresponding aminoacyl-tRNA.

Translation termination is initiated when one of the three stop codons is present in the ribosomal A site, resulting in binding of the Release Factor (RF) proteins. Then, RF1 is removed from the ribosome in a GTP-dependent reaction involving RF3, resulting in the dissociation of the 60S/mRNA complex. However, a peculiar family of selenium containing proteins present in all three domains of life, recode the UGA stop codon into the 21st amino acid, the selenocisteine. The alternative decoding is mediated by a stem-loop structure in the 3′UTR of selenoprotein mRNAs (the SECIS element). See Castellano (2004) for a good review of selenoproteins.

Chemical properties that distinguish different amino acids and post-translational mod-

ifications (i.e. phosporilations, methylation or cleavage) ultimately cause the protein chains to fold up into specific three-dimensional structures that enable them to carry out their specific function.

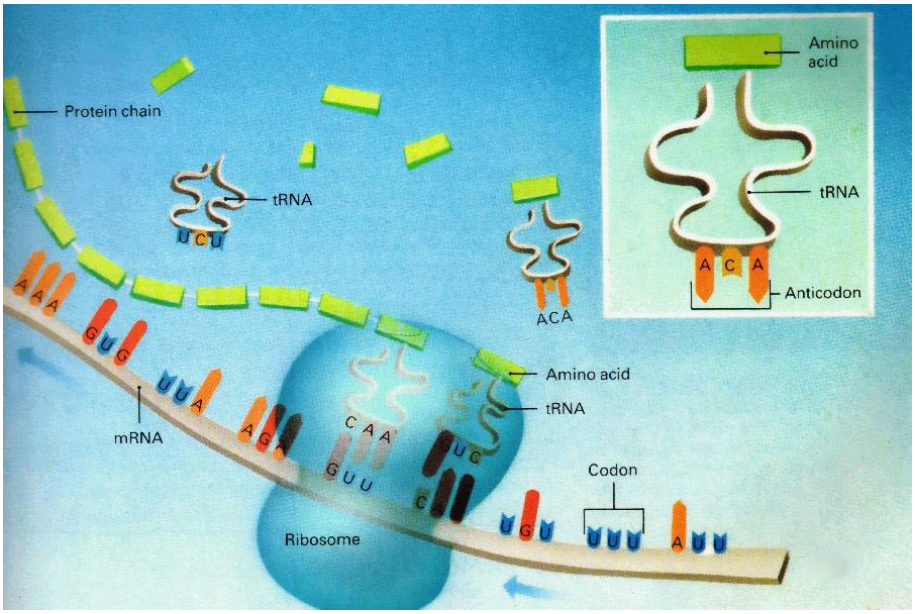


Figure 1.5: Standard protein translation of an mRNA. From *Biology*, Harcout, Brace and Jovanovich (1986).

1.2 Gene prediction

Computational analysis is a major, integral part of genomics, as stated by Galperin and Koonin (2003). From genome shotgun sequence assembly to gene prediction and from sequence comparison to functional protein assignment, including evolutionary studies and building of phylogenetic trees. It would not be an exaggeration to claim that genomics analysis can only be made with computational tools. The way in which genomic information can be directly codified as string of letters make it easy to process, to store and to visualize. A lot of algorithms that had already been developed to analyze and solve string based problems are now being used in the genome analysis field.

The human genome is about 3,000,000,000 base pairs, and many other genomes are already stored in the public databases. Only by using computational methods and statistical models we can try to find out how genes are encoded and try to accurately predict their location in complete genomic sequences.

In the following sections we are going to describe the most common statistical methods to model genes features, and give a brief summary of the previously existent methods

Figure 1.6: Example of genomic sequence. Annotation of the encoded gene is showed in colors. Red boxes correspond to coding regions, the translational start site (ATG) in green, the almost invariable dinucleotides of the splice sites in violet (GT-AG) and the stop codon in blue. From (Blanco, 2000).

for computational gene prediction.

1.2.1 Gene prediction methods

As we have seen in the previous section, a given protein sequence is not usually specified by a continuous DNA sequence, but genes are often split in a number (that may be large) of small coding fragments known as exons, separated by, usually larger, non-coding intervening fragments known as introns. Often, intronic and intergenic DNA is considered to correspond to a large fraction of the genome in higher eukaryotes. In the human genome, for instance, only a very small fraction of the DNA, which can be lower than 2%, corresponds to protein coding exons.

The aim in any gene prediction program is, given a DNA sequence (see Figure 1.6), to find the encoded gene structures and the corresponding amino acid sequences. Indeed, a typical computational eukaryotic gene prediction tool involves the following tasks:

- identification of suitable signals
- prediction of candidate exons defined by the corresponding signals
- assembly of a subset of these exon candidates in a predicted gene structure

The particular implementation of these tasks varies considerably between programs.

Prediction of biological signals

Sequence signals involved in gene prediction are defined as short functional sequence elements that are recognized by the cellular machinery involved in the pathway leading

from DNA to proteins. As has been shown in the previous sections, there are many sequence signals that play a key role in the recognition of genes. However, as gene finding tools focus on predicting coding regions, the four basic signals that define the boundaries of the coding regions of the primary transcript are the most commonly used. These signals correspond to the translational start site (including the Kozak region), the 5′ splice site (also known as donor site), the 3′ splice site (also known as acceptor site) and the translational stop codons.

The main problem in signal recognition is that these motifs are not 100% conserved. There is a high degree of variation in the actual sequences that are recognized by the same cell machinery. Therefore, how can we model this variability? The most widespread approach to model these sequence signals has been position frequency matrices. Senapathy et al. (1990) presented the first quantitative frequency matrices where each matrix element P_{ij} is the frequency of the base i in the position j from a set of aligned splice site sequences.

If Q_{ij} is the background frequency of nucleotide i at position j, then the popular log-odd scoring matrix (known as Position Weight Matrix PWM, Weight Matrix Model WMM, or Position Specific Scoring Matrix PSSM) can be computed as:

$$M_{ij} = log(P_{ij}/Q_{ij}). \tag{1.1}$$

Then, given a sequence of length l the log-likelihood ratio can be computed by summing the coefficients of the log-likelihood matrix corresponding to each nucleotide in each position on the sequence. Usually, a window the length of the matrix is run along the sequence, and the coefficients from the matrix corresponding to each nucleotide in each position on the window sequence are summed. Formally, the score of a matrix M for a site s of length l ($s = s_1, ..., s_l$, and s_k being one of the four nucleotides) is computed as:

$$m_s = \sum_{j=1}^{l} M_{s_j j}. \tag{1.2}$$

If the resulting score is positive the sequence occurs in the signal site more often than in the background model, while if the score is negative the sequence is more likely to be found in the background model. Figure 1.7 shows the frequency matrix and a PWM derived from a set of *D. melanogaster* donor sites.

Although, PWMs are the most common method to model sequence motifs, they have some limitations. For instance PWMs assume that bases at different positions are independent and this is often an oversimplification. Several authors have observed statistically significant dependencies between position within different signals. Certain observed dependencies between donor splice sites positions can be interpreted in terms of the thermodynamics of RNA duplex formation between U1 snRNA and the 5′ splicing region of the pre-mRNA. Similarly, positional dependencies observed in human acceptors sites, appear partially to result simply from the compositional heterogeneity of the human genome, whereas others probably relate to specificity of pyrimidine tract binding proteins.

There are many ways to incorporate base dependencies. One method is to assume that the probability of each base at a given position depends on the base occurring at

a

	-3	-2	-1	0	1	2	3	4	5	6
A	35.14	50.33	12.81	0.00	0.00	58.92	75.69	5.94	11.76	30.91
C	28.01	14.40	8.98	0.00	0.00	1.85	5.94	2.25	12.02	13.74
G	19.29	18.63	67.24	100.00	0.00	35.14	11.36	87.58	7.40	18.49
T	17.57	16.64	10.96	0.00	100.00	4.10	7.00	4.23	68.82	36.86

b

	-3	-2	-1	0	1	2	3	4	5	6
A	26.25	28.41	28.11	0.00	0.00	20.86	24.29	27.98	24.87	27.14
C	21.93	22.64	19.47	0.00	0.00	20.57	24.55	22.50	24.85	22.36
G	23.40	21.92	21.32	100.00	0.00	30.17	23.86	23.25	21.88	25.10
T	28.34	26.97	31.05	0.00	100.00	28.37	27.24	26.20	28.31	25.28

c

	-3	-2	-1	0	1	2	3	4	5	6
A	0.292	0.572	-0.786	.	.	1.038	1.137	-1.550	-0.749	0.130
C	0.245	-0.452	-0.774	.	.	-2.409	-1.419	-2.303	-0.726	-0.487
G	-0.193	-0.163	1.149	.	.	0.152	-0.742	1.326	-1.084	-0.306
T	-0.478	-0.483	-1.041	.	.	-1.934	-1.359	-1.824	0.888	0.377
	A/C	A	G	G	T	A/G	A	G	T	G

Figure 1.7: Frequency matrices and position weight matrix derived from a set of canonical donor splice sites. Negative positions correspond to the bases upstream of the splice site boundaries. **a)** Frequency matrix derived from a set of canonical donor splice sites. **b)** Background frequency matrix derived from a set of sequences with the a conserved GT but not annotated as fucntional donor sites. **c)** Position weight matrix computed from the frequency matrix derived from the real donor sites versus the background frequency matrix. The consensus sequence is showed below.

the one or more of the previous positions (the so-called Position Weight Array PWA, Markov dependence or Weight Array Model WAM, Zhang and Marr (1993)). Another method is to apply a decision tree (so-called Maximal Dependence Decomposition MDD, Burge and Karlin (1997)) to partition the training data into subsets so that splice site bases within each subset are approximately independent and can, hence, be modeled by separate PWMs.

Recently described methods, as multilayer neural networks (Reese et al., 1997) or inclusion-driven learned Bayesian Networks (idlBNs) (Castelo and Guigó, 2004), model significant dependencies between all possible bases. More dependencies, however, result in more parameters to be estimated and require much more data to generate the models. These more complex models typically yield significant, but not dramatic, improvements in splice site discrimination over the simpler models which assume only dependence between adjacent positions. The increase of efficiency is higher in isolation than in the context of an integrating gene finding method, where other information (like coding statistics) also helps, indirectly, in the definition of the exon boundaries (Burge and Karlin, 1998).

Prediction of exons

Once all the signals are predicted, all the putative exons can be built. Typically, gene prediction programs redefine the term exon to refer only to the coding fraction of the exons, and classify them as: initial (limited by a translational start site and a donor site), internal (limited by an acceptor and a donor site), terminal (limited by an acceptor and a stop codon) and single exon genes (limited by a translational start site and a stop codon). See Zhang (2002) for a more realistic description of all the possible types of exons including UTR exons and mixed coding and non-coding regions.

To discriminate coding regions from non coding regions a large number of content measures have been developed (Fickett and Tung (1992), Gelfand (1995) and Guigó (1999)). Such content measures, also known as coding statistics, can be defined as functions that compute a real number that indicates the likelihood of a given DNA sequence to code for a protein. Protein coding regions exhibit characteristic sequence composition which is absent in non-coding regions. This bias is mainly due to coding restrictions: the specific amino acid usage to build proteins and the unequal usage of the synonymous codons (see Table 1.1). Fickett (1982) also showed that coding regions have asymmetries and periodicities that help to distinguish them from non-coding sequences.

Among all the different coding measures, codon position dependent 5th order Markov models (Borodovsky and McIninch, 1993) appear to offer the maximum discriminative power (Guigó, 1998) and are at the core of most popular gene finders today. In this model, the conditional probability of the identity of the next nucleotide depends on the identities of previous five bases. This relatively complicated model incorporates a combination of biases related to amino acid usage, codon usage, di-amino acid and dicodon usage as well as other underlying factors.

Coding measures are usually combined with the scores of the exon defining signals to obtain a final exon score. There are a number of ways in which these scores can be combined. If computed as log-odds, they can be simply summed up under the assumption of independence.

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

Table 1.1: The human codon usage and codon preference table. Published online at `http://bioinformatics.weizmann.ac.il/databases/codon`. For each codon, the table displays the frequency of usage of each codon (per thousand) in human coding regions (first column) and the relative frequency of each codon among synonymous codons (second column).

Assembly of putative exons

Predicted exons need to be assembled into genes. This assembly must conform to a number of intrinsic biological constraints such as non-overlap between assembled exons and the maintenance of an open reading frame (ORF) among them.

The main difficulty in exon assembly is the combinatorial explosion problem: the number of ways N candidate exons may be combined grows exponentially with N. To address this problem a number of methods based on dynamic programming techniques have been developed. In dynamic programming, the solution to a general problem is obtained by the recursive solution of smaller versions of the problem. In the "optimal exon assembly" problem, dynamic programming allows us to find the solution efficiently, without having to enumerate all exon assembly possibilities (Gelfand and Roytberg, 1993). Algorithms running in quadratic time (in time proportional to the square of the number of predicted exons, $O(N^2)$) were used in `geneparser` (Snyder and Stormo, 1993), `grailII` (Xu et al., 1994) and `fgenes` (Solovyev et al., 1995), among other programs. Guigó (1998) developed a more efficient algorithm running in linear time (that is in time proportional to the number of predicted exons, $O(N)$). At the core of the recently developed `gaze` (Howe et al., 2002), a program that assembles data obtained from external sources of gene predictions and experimental evidence, there is also a chaining algorithm that runs effectively in linear time.

A revolution in gene prediction was the application of Generalized Hidden Markov Models (GHMMs). This probability model was first developed in the speech-recognition field and later applied to protein and DNA sequence pattern recognition and was initially implemented in the gene prediction field in the `genie` algorithm (Kulp et al., 1996).

In a GHMM approach, different types of gene structure components (such as exon or intron) are characterized with states (as shown in Figure 1.8). A gene model is generated by a state machine: starting from 5′ to 3′, each base-pair is generated by a "emission probability" conditioned on the current state and surrounding sequences and transition from one state to another is governed by a "transition probability" which obeys all the con-

straints (for example an intron can only follow an exon, reading frames of two adjacent exons must be compatible, etc.). All the parameters of the "emission probabilities" and "transition probabilities" are learned (pre-computed) from some training data set. Since the states are unknown ("hidden"), an efficient dynamic programming algorithm (called the Viterbi algorithm) may be used to select the best set of consecutive states (called a "parse"), which has the highest overall probability compared with any other possible parse of the given genomic sequence (see Rabinier (1989) for a tutorial on GHMMs).

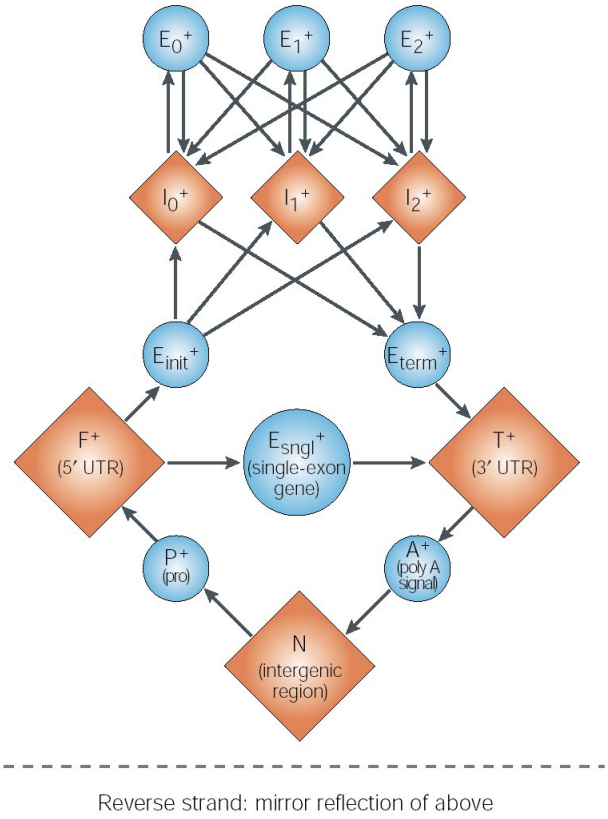


Figure 1.8: Schema of the states and transitions in `genscan` GHMM. Each circle or square represents a functional unit (a state) of a gene. Arrows represent the transition probability from one state to another. *E* correponds to exon, *I* to intron and *pro* to promoter. Adapted from Burge and Karlin (1997).

1.2.2 Ab initio gene prediction

Computational gene finding is not a brand new field and a large body of literature has accumulated during the last 25 years. Early studies by Shepherd (1981), Fickett (1982) and Staden and McLachlan (1982) showed that statistical measures related to biases in amino

acid and codon usage could be used to approximately identify protein coding regions in genomic sequences. Based on these differences, the first generation of gene predictions programs, designed to identify approximate locations of coding regions in genomic DNA were developed. The most widely known of such kind of programs were probably `testcode` (based on Fickett (1982)) and `grail` (Uberbacher and Mural, 1991). These programs were able to identify coding regions of sufficient length (100-200 bp) with fairly high reliability, but did not accurately predict exon locations.

In order to predict exon boundaries, a new generation of algorithms were developed. A second generation of programs, such `sorfind` (Hutchinson and Hayden, 1992), `grailII` (Xu et al., 1994) and `xpound` (Thomas and Skolnick, 1994), use a combination of splice signal and coding region identification techniques to predict "spliceable open reading frames" (potential sets of exons), but do not attempt to assemble predicted exons into complete genes. A third generation of programs attempt the more difficult task of predicting complete gene structures: sets of exons which can be assembled into translatable coding sequences. The earliest examples of such integrated gene finding algorithms were probably the `genemodeler` program (Fields and Soderlund, 1990) for prediction of genes in *C. elegans* and the method of Gelfand (1990) for mammalian sequences. Subsequently, there has been a mini-boom of interest in development of such methods, and a wide variety of programs have appeared, including (but not limited): `geneid` (Guigó et al., 1992), which used a hierarchical rule based structure; `geneparser` (Snyder and Stormo, 1993), which scored all subintervals in a sequence for content statistics and splice site signals weighted by a neural network and chained by dynamic programing; `genemark` (Borodovsky and McIninch, 1993) which combined the specific Markov models of coding and non-coding region together with Bayes' decision making function; `genlang` (Dong and Searls, 1994), which treated the problem by linguistic methods describing a grammar and parser for eukaryotic protein-encoding genes; and `fgenes` (Solovyev et al., 1994) which used a discriminant analysis for identification of splice sites, exons and promoter elements.

At the end of last decade, the introduction of the GHMMs produced a new generation of gene prediction programs. GHMMs, as discussed in the previous section, have some advantages over the previous approaches. The main advantage is that all the parameters of the model are probabilities and that, given a set of curated sequences and defined states, the Viterbi algorithm can be used to compute the set of optimal parameters. A great variety of programs appeared simultaneously exploring the capabilities of GHMMs: `genie` (Kulp et al., 1996), `hmmgene` (Krogh, 1997), `veil` (Henderson et al., 1997), `genscan` (Burge and Karlin, 1997) and the GHMMs version of `genemark` (`genemark.hmm`, Lukashin and Borodovski (1998)) and `fgenes` (`fgenesh`, Salamov and Solovyev (2000)).

Of the gene prediction tools that were released during this period, `genscan` clearly outperformed all the others (at least with regards to human gene prediction). Novel features included in `genscan` were: the capacity to predict multiple genes in a sequence, to deal with partial as well as complete genes, and to predict consistent sets of genes occurring on both DNA strands; the use of distinct, explicit, empirically derived sets of model parameters to capture differences in gene structure and composition between different C+G compositional regions; and statistical models of donor (using MDDs) and acceptor (using PWAs) splice sites which capture potentially important dependencies between signal positions. Significant improvements in the accuracy of prediction have been observed

for genscan over existing programs at that time.

genscan is still considered the standard gene prediction program (at least for human) and it is used in most of the genome annotation pipelines like ENSEMBL and the NCBI genome resources.

1.2.3 Genome comparison gene prediction

With the availability of many genomes from different species, a number of strategies have been developed to use genome comparisons to predict genes. The rationale behind comparative genomic methods is that functional regions, protein coding among them, are more conserved than non-coding ones between genome sequences from different organisms (see Figure 1.9). This characteristic conservation can be used to identify protein coding exons in the sequences. The approach taken by different programs to exploit this idea differ notably.

In one such approach (Blayo et al., 2002; Pedersen and Scharl, 2002), the problem is stated as a generalization of pairwise sequence alignment: given two genomic sequences coding for homologous genes, the goal is to obtain the predicted exonic structure in each sequence maximizing the score of the alignment of the resulting amino acid sequences. Both, Blayo et al. (2002) and Pedersen and Scharl (2002) solve the problem through a complex extension of the classical dynamic programming algorithm for sequence alignment. Although very appropriate for short sequences, in practice, the time and memory requirements of this algorithm may limit its utility for very large genomic sequences. Moreover, although the approach theoretically guarantees to produce the optimal amino acid sequence alignment, the fact that sequence conservation may also occur in regions other than protein coding, could lead to over prediction of coding regions, in particular when comparing large genomic sequences from homologous sequences from closely related species.

To overcome this limitation, the programs doublescan (Meyer and Durbin, 2002) and slam (Alexandersson et al., 2003) rely on more sophisticated models of coding and non-coding DNA and splice signals, in addition of sequence similarity. Since sequence alignment can be solved with Pair Hidden Markov Models (PHMMs, Durbin et al., 1998) and GHMMs have been proved very useful to model the characteristics of eukaryotic genes (Burge and Karlin, 1997), slam and doublescan are built upon the so-called Generalized Pair HMMs. In these, gene prediction is not the result of the sequence alignment, as in the programs above, but both gene prediction and sequence alignment are obtained simultaneously.

A third class of programs adopt a more heuristic approach, and separate clearly gene prediction from sequence alignment. The programs rosetta (Batzoglou et al., 2000), sgp1 (from Syntenic Gene Prediction, Wiehe et al., 2001), and cem (from the Conserved Exon Method, Bafna and Huson, 2000) are representative of this approach. All these programs start by aligning two syntenic regions (specifically human and mouse in rosetta, and cem; less species specific in sgp1), using some alignment tool (the glass program, specifically developed in the case of rosetta or generic ones, such as tblastx, or sim96 in the case of cem and sgp1), and then predict gene structures in which the exons are compatible with the alignment. This compatibility often requires conservation of exonic structure of the homologous genes encoded in the anonymous syntenic

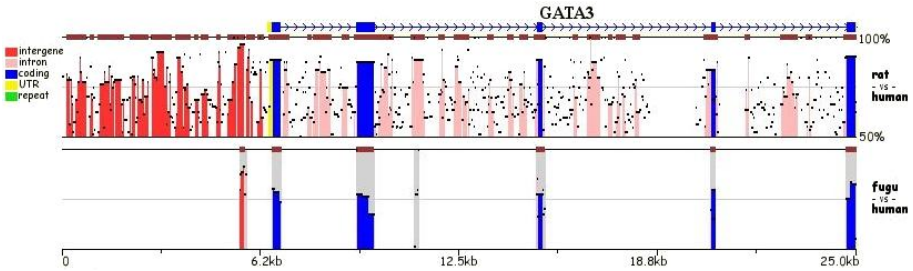


Figure 1.9: A plot of sequence conservation across the *gata*3 gene region in human, rat and *Takifugu*with the zpicture program. From `http://zpicture.dcode.org/zPicture.php?id=example&numseq=3`.

regions. Although conservation of exonic structure is an almost universal feature of orthologous human/mouse genes (Mouse Genome Sequencing Consortium, 2002), it does not necessarily occur when comparing genomic sequences of homologous genes from other species.

As the number of genome sequences of species at different evolutionary distances increases, methods to predict genes based on the comparative analysis of multiple genomes (and not only of two species) look promising. For instance, Dewey et al. (2004) combine pairwise predictions from `slam` in the human, mouse and rat genomes to simultaneously predict genes with conserved exonic structure in all three species. In the so-called Phylogenetic Hidden Markov Models (phylo-HMMs) or Evolutionary Hidden Markov Models (EHMMs), a gene prediction Hidden Markov Model is combined with a set of evolutionary models, based on phylogenetic trees. Phylo-HMMs take into account that the rate (and type) of evolutionary events differ in protein-coding and non-coding regions. Recently, phylo-HMMs have been applied to gene prediction with encouraging results (Pedersen and Hein, 2003; Siepel and Haussler, 2004).

1.2.4 Gene prediction accuracy

The accuracy of gene prediction programs is usually measured in controlled data sets. To evaluate the accuracy of a gene prediction program, the gene structure predicted by the program is compared with the structure of the actual gene encoded in the sequence. The accuracy can be evaluated at different levels of resolution. Typically, these are the nucleotide, exon, and gene levels. These three levels offer complementary views of the accuracy of the program. At each level, there are two basic measures: sensitivity and specificity. Briefly, sensitivity (Sn) is the proportion of real elements (coding nucleotides, exons or genes) that have been correctly predicted, while specificity (Sp) is the proportion of predicted elements that are correct. More specifically, if TP is the total number of coding elements correctly predicted, TN, the number of correctly predicted non-coding elements, FP the number of non-coding elements predicted coding, and FN the number of coding elements predicted as non-coding (see Figure 1.10). Then, in the gene finding literature, Sn is defined as:

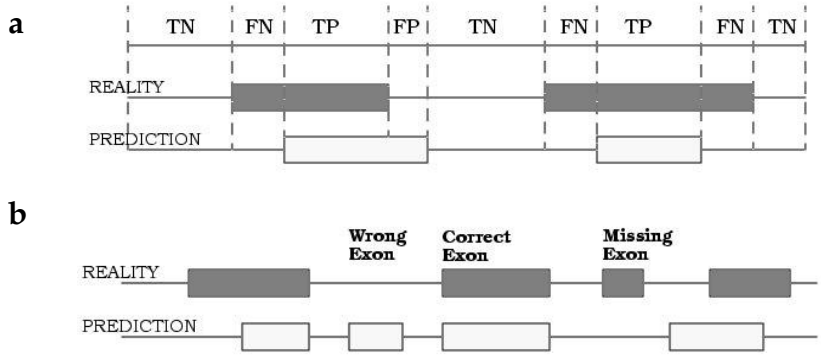


Figure 1.10: Schema of the measures used to determine gene prediction accuracy. **a)** Definition of the TN true negatives, FN false negatives, TP true positives, and FP false positives, when the evaluation is perform at base level. **b)** Examples of perfect macth, and missing and wrong exons. From Burset and Guigó (1996).

$$Sn = \frac{TP}{TP + FN} \tag{1.3}$$

and Sp as:

$$Sp = \frac{TP}{TP + FP} \tag{1.4}$$

Both, Sn and Sp, take values from 0 to 1, with perfect prediction when both measures are equal to 1. Neither Sn nor Sp alone constitute good measures of global accuracy, since high sensitivity can be reached with low specificity and vice versa. It is desirable to use a single measure for accuracy. In gene finding literature, the preferred such measure at the nucleotide level is the Correlation Coefficient, which is defined as:

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{1.5}$$

and ranges from -1 to 1, with 1 corresponding to a perfect prediction, and -1 to a prediction in which each coding nucleotide is predicted as non-coding and vice versa.

At exon level, these measures determine if predictions correspond to real exons, with the exon boundaries perfectly predicted (see Figure 1.10). The prediction is considered incorrect if only a single base does not correspond to the coordinates of the real exon. Therefore, Sn at exon level measures the proportion of actual exons that have been perfectly predicted, and Sp measures the proportion of predicted exons that correspond to actual exons. The average exon prediction accuracy $SnSp$ is computed as:

$$SnSp = \frac{Sn + Sp}{2} \tag{1.6}$$

Apart from Sn, Sp and $SnSp$, two extra measures have been used to determine the accuracy at exon level: the missed exons (ME) and the wrong exons (WE). ME measures how frequently a predictor completely failed to identify exons (no prediction overlap at all) whereas WE identifies the ratio of exons that do not overlap with any exon of the standard set. At gene level Sn and Sp measure if a predictor is able to correctly identify and assemble all of the exons of a gene. For a prediction to be counted as TP, all coding exons must be identified, every intron-exon boundary must be exactly correct, and all the exons must be included in the proper gene. In addition, missed genes (MG) and wrong genes (WG) can also be computed in the same way as at the exon level.

The large amount of gene finding programs that have been described in the previous sections raises the obvious question of whether the gene finding problem has perhaps already been solved. This question was repetitively answered negatively by different systematic comparisons of available integrated gene finding methods.

Table 1.2 reproduces the results from the benchmark by Burset and Guigó (1996), one of the first systematic evaluations of gene finders. These authors evaluated seven programs, using a set of 570 vertebrate single gene genomic sequences deposited in GenBank after January 1993. This was done to minimize the overlap between this test set and the sets of sequences which the programs had been trained on. The average CC for the programs analyzed ranged from 0.65 to 0.80 at the nucleotide level, while the $SnSp$ at exon level ranged from 0.37 to 0.64.

Recently, a new independent comparative analysis of seven gene prediction programs have been published (Rogic et al., 2001). The programs were again tested in a set of 195 single gene sequences from human and rodent species. In order to avoid overlap with the training sets of the programs, only sequences were selected that had been entered in GenBank, after the programs were developed and trained. Table 1.3 shows the accuracy measures averaged over the set of sequences effectively analyzed for each of the tested programs.

The programs tested by Rogic et al. (2001) showed substantially higher accuracy than the programs tested by Burset and Guigó (1996): the average CC at the nucleotide level ranged from 0.66 to 0.91, while the average exon prediction accuracy ranged from 0.43 to 0.76. This illustrates the significant advances in computational gene finding that were achived during the nineties.

The evaluations by Burset and Guigó (1996), Rogic et al. (2001), and others suffered from the same limitation: gene finders were tested in controlled data sets made of short genomic sequences encoding a single gene with a simple gene structure. These datasets are not representative of the genome sequences being currently produced: large sequences of low coding density, encoding several genes and/or incomplete genes, with complex gene structures.

	Base level			Exon level				
	Sn	Sp	CC	Sn	Sp	SnSp	WE	ME
fgenes	0.77	0.88	0.80	0.61	0.64	0.64	0.15	0.12
geneparser2	0.66	0.79	0.65	0.35	0.40	0.37	0.29	0.17
genlang	0.72	0.79	0.71	0.51	0.52	0.52	0.21	0.22
grail 2	0.72	0.87	0.76	0.36	0.43	0.40	0.25	0.11
sorfind	0.71	0.85	0.72	0.42	0.47	0.45	0.24	0.14
xpound	0.61	0.87	0.69	0.15	0.18	0.17	0.33	0.13
geneid+	0.91	0.91	0.88	0.73	0.70	0.71	0.07	0.13
geneparser3	0.86	0.91	0.85	0.56	0.58	0.57	0.14	0.09

Table 1.2: Evaluation of the different gene finding tools. The evaluation is divided into nucleotide level and exon level. From Burset and Guigó (1996).

	Base level			Exon level				
	Sn	Sp	CC	Sn	Sp	SnSp	WE	ME
fgenesh	0.86	0.88	0.83	0.67	0.67	0.67	0.12	0.09
genemark.hmm	0.87	0.89	0.83	0.53	0.54	0.54	0.13	0.11
genie	0.91	0.90	0.88	0.71	0.70	0.71	0.19	0.11
genscan	0.95	0.90	0.91	0.70	0.70	0.70	0.08	0.09
hmmgene	0.93	0.93	0.91	0.76	0.77	0.76	0.12	0.07
morgan	0.75	0.74	0.69	0.46	0.41	0.43	0.20	0.28
mzef	0.70	0.73	0.66	0.58	0.59	0.59	0.32	0.23

Table 1.3: Evaluation of the different gene finding tools. The evaluation is divided into nucleotide level and exon level. From Rogic et al. (2001).

1.3 Automatic genome annotation pipelines: ENSEMBL

To annotate a genome is, in short, to identify (find the start and end coordinates along a DNA sequence) the key features of the genome (i.e. genes, promoter regions, polymorphisms). Usually, we refer to an annotation pipeline as an automatic (computational) or semi-automatic (with human intervention) process in which these features are predicted, somehow assessed (by computational or experimental means) and this information gathered in a comprehensible way. This is achieved by the combination of several computational programs which analyze different aspects of the genomic sequence. This process may also include a user-friendly display interface, which makes this biological information available to the whole scientific community.

There are three main systems that annotate and display genome information: ENSEMBL (`http://www.ensembl.org`), the University of California (Santa Cruz) genome browser system (UCSC browser, `http://genome.cse.ucsc.edu/`) and the National Center of Biotechnology Information genome resources (NCBI browser, `http://www.ncbi.nlm.nih.gov`). ENSEMBL is considered to generate the most reliable set of genome annotations and many genome projects consider it as a standard reference.

The ENSEMBL project was conceived in response to the acceleration of the public effort to sequence the human genome in 1999. At that time, it was clear that if the annotation of the draft sequence was to be available in a reasonable amount of time, it had to be automatically generated to deal with the new genomes to come and with the subsequent releases.

The initial stage of the automated genome annotation in ENSEMBL starts with running a set of analysis tools. It includes `repeatmasker` (Smit and Green, 1999), `genscan`, `tRNAscan` (Lowe and Eddy, 1997), `eponine` (Down and Hubbard, 2002) and homology searches using `blast`. The results from this initial analysis are combined in a complex automatic process to generate the final annotation.

The ENSEMBL gene-build process is based on genomic information coming from four different sources: proteins and mRNAs from the corresponding species, proteins and mRNAs from other species, expressed sequence tags (ESTs) and *ab initio* gene predictions supported by experimental data. The complete pipeline is described in depth in Curwen et al. (2004) and can be briefly summarized as follows (see Figure 1.11):

- Proteins and mRNAs from the species whose genome is being annotated are mapped in the genome to create transcript models. First, proteins of the genome of interest are aligned against the entire genome using `pmacth` (Durbin, unpubl.). The second stage is to realign the proteins in the corresponding region with a more accurate (and time consuming) program (`genewise` (Birney et al., 2004b) in the case of proteins and `est_genome` (Mott, 1997) for mRNAs). Protein and mRNA based transcripts are combined to obtain transcripts with untranslated region information.

- Proteins and mRNAs from other species are then used to locate the transcripts which have not been found previously. The same two-step approach is used but, less restricted parameters are used to allow some degree of divergence.

- The ENSEMBL EST gene build process involves three steps. First, ESTs from the species of interest are aligned against the entire genome using `exonerate` (Slater,

unpubl.). The second stage is to realign the ESTs in a smaller region with the more accurate program `est_genome`. In the third step the aligned ESTs are used to build all compatible gene structures using the `clustermerge` (Eyras et al., 2004) algorithm.

- *Ab initio* predicted genes are compared against different DNA and protein sequences databases using `blast`. Using this information putative transcripts are generated in the following way: adjacent exon pairs are built if they are supported by `blast` evidence in a consistence way (neither overlapping nor having a excessive gap between them). Exon pairs are then recursively linked into transcripts which can be clustered together.

After the gene-building process all predictions are gathered and labeled with the corresponding identification (consistent among different releases).

As we have seen, all ENSEMBL predictions are at least partially based on preexisting evidence of transcription or similarity to known proteins. Thus, the ENSEMBL pipeline is biased to produce a set with high specificity at the expense of sensitivity: they prefer to miss a few features than heavily overpredict genes. As extensively discussed in Birney et al. (2004a) there are two reasons that lead them to follow these criteria. First, there are already several programs that generate high sensitivity at the expense of specificity and ENSEMBL already provides the results of some of these tools through their web site. Second, they considered that specific data sets are more useful for researchers in order to assure a high ratio of success in experimental approaches to study or to characterize any of the predicted genes.

1.4 Experimental verification of gene predictions

Once we have a set of gene predictions, it would be desirable to have a systematic way to validate experimentally whether they correspond to actual genes. The most intuitively way to determine if predicted genes are functional would be to find the encoded proteins expressed in the corresponding organism. Very promising advances have been achieved in the determination of genomic coding regions with the analysis of two dimensional protein gels and subsequent mass spectrometry (Arthur and Wilkins, 2004). However these techniques are still in a very early stage of development for whole genome approaches.

Other evidence of the expression of a gene, is the evidence of transcription of the genomic region where it is encoded. Although it can not be claimed that the transcript is translated into the predicted protein, translation and splicing are considered strong evidence of functionality. There are two main techniques to identify and characterize expressed mRNAs: microarrays and RT-PCR amplification.

1.4.1 Microarrays

DNA microarray or DNA chip technology allows the monitoring of the expression of thousand of genes at the same time. Microarrays are rigid supports on which oligonucleotide probes have been synthesized *in situ* or deposited by high-speed robotic printing. Transcripts from two different sources or cell conditions are obtained and usually

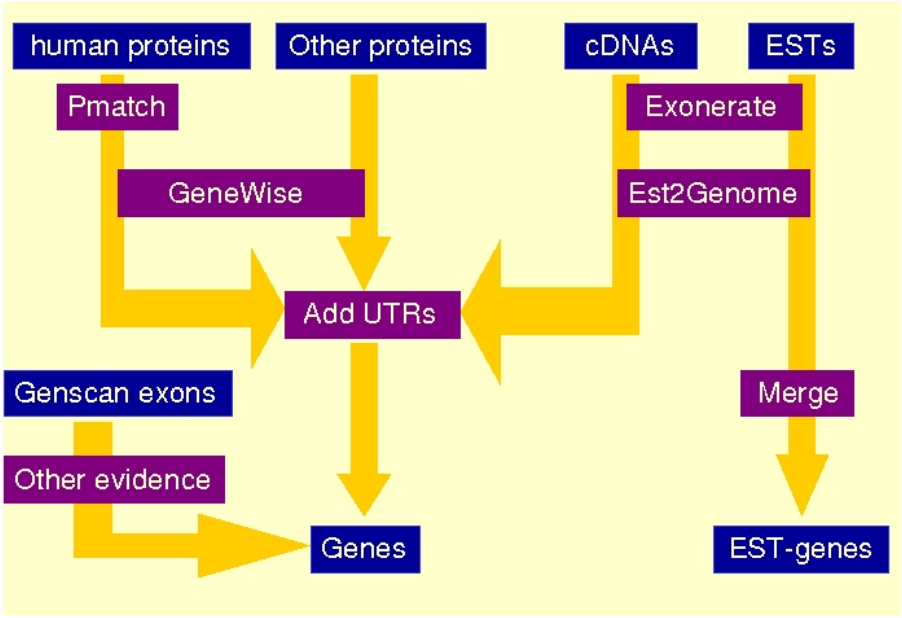


Figure 1.11: Schema of the automatic annotation pipeline used by ENSEMBL. Figure kindly provided by Eduardo Eyras.

converted to the more stable complementary DNA and labeled with two different fluorescent markers. Labeled transcripts are mixed with the oligonucleotide probes that are attached to the surface of the substrate of the microarray. After hybridization, spots are washed to remove unhybridized transcripts. Then, the microarray is scanned using two different lasers, corresponding to the excitation wavelength of the markers. The fluorescence signal from each transcript population is evaluated independently and used to calculate the expression ratio.

Microarray based methods have recently been applied to verify novel gene predictions. For example, Penn et al. (2000) tested a collection of ORFs predicted by different gene finding programs in the draft sequence of the human genome and Shoemaker et al. (2001), tested the expression of all annotated exons predicted by `genscan` in human chromosome 22 under 69 different conditions. In the experiment showed in Shoemaker et al. (2001), two 60-mer oligonucleotides were designed based on each predicted exon and printed on a single array. This array was hybridized with 69 pairs of RNA samples using two colors hybridization technique (see Figure 1.12). New genes were verified as groups of co-expressed exons that are located next to each other in the genome. Although microarrays offer an attractive approach for large-scale monitoring of mRNA levels, the approaches described in these studies cannot directly determine whether two exons form part of the same transcript or are part of two coexpressed genes, relying on co-expression to make such inferences.

Recent experiments have illustrated the principle that microarrays can monitor splicing events, using probes positioned at exon-exon junctions (Johnson et al., 2003). De-

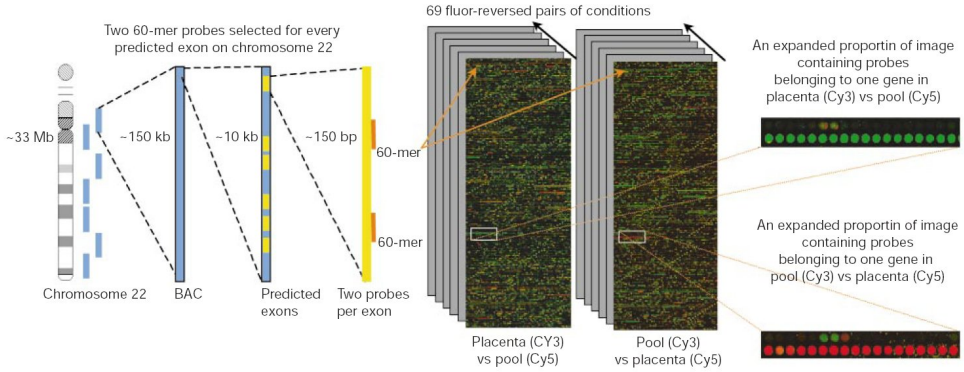


Figure 1.12: Design and fabrication of exon arrays for the predicted exons on human chromosome 22. From Shoemaker et al. (2001).

tection of expression using "junction arrays" is limited in several ways. Junction arrays cannot determine whether two splicing events in one tissue are present in the same or separate transcripts. Detection also requires differential expression; if two isoforms are present in the same proportion in every tissue, no signal will be observed. Finally, cross hybridization could cause false positives when sequence-similar genes have strong tissue specific regulation. The resolution and sensitivity of this approach could be improved by adding probes in exons.

1.4.2 RT-PCR amplification

RT-PCR (reverse-transcriptase polymerase chain-reaction) allows the amplification of small amounts of RNA fragments and is the most sensitive technique for mRNA detection and quantification currently available. Compared to the two other commonly used techniques for quantifying mRNA levels, Northern blot analysis and RNAse protection assay, RT-PCR can be used to quantify mRNA levels from much smaller samples. In fact, this technique is sensitive enough to enable quantification of RNA from a single cell.

Figure 1.13 shows the schema of the amplification of a target sequence using RT-PCR. First, the mRNA must be isolated from tissue or cells and made accessible to the primers. To generate the cDNA using the enzyme reverse transcriptase (RT), the primer must be attached to the mRNA target. Then, the first strand of the cDNA is synthesized producing a hybrid molecule that consists of the mRNA template and the complementary DNA strand. In the next step the template strand of RNA is removed by treatment with RNAse II. What follows is a typical PCR amplification. The second primer is bound to the template cDNA and the Taq polymerase adds the complementary nucleotides. The resulting product is a double stranded cDNA. The three step process of denaturation, primer binding and Taq extension is repeated to yield a detectable PCR product, the product can be visualized on ethidium bromide stained agarose gel following electrophoresis.

In some cases RT-PCR can produce false positives due to amplification of genomic

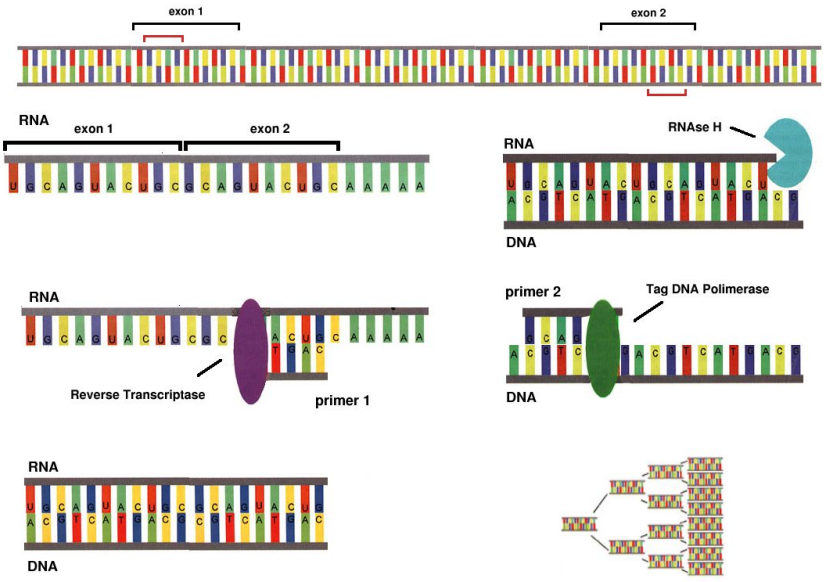


Figure 1.13: Schema RT-PCR amplification process. The process is shown from top to bottom and from left to rigth. Adapted from `http://ccm.ucdavis.edu/cpl/Techupdates/TechUpdates.htm`.

DNA instead of RNA. In the case of multi-exonic gene structure validation, primers are located in exons that, in most cases, are separated more than the number of bases that the reverse transcriptase is able to transcribe in a row. Therefore, only after the splicing events that join the two exons, the primers are at the optimal distance for the amplification. If splicing does not occurs, the probes are too far away from each other, so that the reverse transcriptase stops before reaching the region where the second primer binds, and thus, the fragment can not be amplified.

Amplified fragments are usually sequenced to confirm that they correspond to the predicted transcripts and to ensure that introns have been removed.

RT-PCR experiments have also been used for large scale validation of gene predictions. In Das et al. (2001) gene predictions on chromosome 22 were validated using primers designed to amplify a pair of adjacent exons. From the results, they infer that approximately between 13% and 27% of the predictions of `genscan` in the chromosome 22 that do not overlap previously annotated genes are considered to be positive.

Objectives

`geneid` (Guigó et al., 1992) was one of the first programs to predict full exonic structures of vertebrate genes in anonymous DNA sequences. However, since the original `geneid` was released, there had been substantial developments in the field of computational gene identification, and it had become clearly inferior to the other existing tools.

The goal of this thesis, was to improve `geneid` prediction accuracy, and make it useful for the new genomes that were going to come. Therefore, the main objectives of this dissertation were the followings:

- To develop and test a generic parameter file structure for the new version of `geneid` including the most appropriate recognition models. The parameter file should have a simple and intuitive interpretation and should be easily estimated from any available set of genes.
- To analyze the signals and intrinsic properties of gene codification in eukaryotes. Check which of the current statistical models better fit each genomic feature and try to develop a more general biological model of the complete process.
- To build sets of reliable annotated genomic sequences for different species and infer from them parameter files for `geneid`.
- To infer evolutionary relations and the evolution of gene codification from the previous generated sets of sequences.
- To develop a method to incorporate genomic comparative information to `geneid` prediction framework.
- To provide and distribute both, predicted genes and the bioinformatic tools to the research community.

Many of the goals listed above have been achieved by the current implementation of `geneid` and its extension, which uses comparative information, `sgp2`.

The parameter file was designed to incorporate several types of information. Depending on the amount of available data for each species and the nature of the signal every type of site could be represented with a position weight array of different order. As coding statistics `geneid` allows the use of any order of Markov chain, depending again on

the amount of available data. Moreover, the new parameter file, can have a complete set of parameters for different C+G content context.

`geneid` and `sgp2` accuracy have been tested in different sets of sequences, showing an accuracy superior to the existing tools, being both specially more specific than other existing programs. On the other hand, certain features remain difficult to predict including very small exons and the exact boundaries of genes. More general challenges in the gene prediction field are pointed out in the *Discussion* section.

Some of the work presented in this dissertation has been done in collaboration with international genome sequencing consortiums. These collaborations gave me the opportunity to meet and work with specialist from all over the world, and made our work very relevant. These collaborations had put a lot of pressure on us and a lot of effort have been invested in the genomic annotation projects. The annotation of recently sequenced genomes, however, has been very fruitful allowing us to test and adapt our gene prediction tools to the real needs of the genomic annotation projects.

On the other hand, this effort was detrimental to some of the initial objectives. For instance, the biological approach to the definition of the splice sites and the building of more realistic models has been impossible to achieve during the realization of this thesis. The comparative analysis of the signals and properties of protein coding genes across the evolution is in a very preliminary stage. Although a lot of information has been gathered building the training sets for `geneid`, we did not have enough time to analyze this data in depth.

For the last objective, the dissemination of data, all the programs, data sets and gene predictions have been made available with no restriction through our own web service and ENSEMBL and the UCSC genome browser systems serve our predictions thought their web interface browsers.

Ab initio gene finding: geneid

This chapter describes the basic `geneid` architecture, the statistical models and the parameters included in the current distribution. The first parameter set for the new `geneid` version was obtained while the re-programing in C of the first version was still in process. The motivation to start working on *D. melanogaster* was the announcement by the Berkeley Drosophila Genome Project of an experiment to determine the state of the art of gene prediction tools in which any bioinformatic group could participate. Our group decided to take part in this assessment and to develop a parameter file for *Drosophila melanogaster*. The results of the assessment are briefly commented. After that, parameter files for several species have been built. A short description of the "training" process and some observed properties of gene codification in different species are presented.

3.1 geneid architecture and parameter file

`geneid` is a program that predicts genes in anonymous genomic sequences designed following a simple hierarchical structure (see Figure 3.1). First, splice sites and translational start and stop codons are predicted and scored along the sequence. Next, potential exons are built from the previously predicted sites and scored, taking into account the score of the sites and the coding sequence model. Finally, from the set of predicted exons, the gene structure maximizing the sum of the score of its exons is assembled using a dynamic programming algorithm (`genamic`, Guigó (1998)).

In most gene prediction programs, there is a clear separation between the gene model itself and the parameters of the model. Typically, the parameters of the gene model define the characteristic of the sequence signals involved in gene specification (i.e. PWMs for the splice sites), the codon bias characteristic of coding exons (i.e. hexamer counts or Markov Models for coding regions), and the relation between the exons when assembled into gene models (i.e. intron and exons length distributions, transition probabilities in GHMMs, etc.). These parameters are estimated from a set of annotated genomic sequence from the species of interest.

The `geneid` parameter file contains the description of the probabilistic models (computed as log-likelihood ratios) in a comprehensible data structure. The file is text-based and includes comments to clarify and to differentiate each defined structure. The definition of each feature has some flexibility allowing different types of models depending

on the amount of training data available. In what follows, the different models and the process of genes prediction in geneid are described. Although some of this information is also included in the paper presented in section 3.2.2, the original models have been extended. For instance, instead of the initial PWMs, the current version allows PWAs (a generalization of the classical position weight matrices) for the detection of the signals. As coding measure, instead of a fixed 5th order Markov Model, a Markov Model of any order can be used.

3.1.1 Site definition

geneid uses PWAs (in which every position contains a Markov chain of order k) to predict acceptor and donor splice sites and start codons. From a collection of annotated sequences containing the same signal, a probability matrix P is derived for the positions around the characteristic motif (i.e. GT for donor sites). Thus, $P^j(x_{k+1}|x_1 \dots x_k)$ is the probability of observing the nucleotide x_{k+1} after the oligonucleotide $x_1 \dots x_k$ at position j in an actual site. A false site is considered to be any sequence that contains the characteristic motif but has not been annotated as a functional site. Therefore, from a collection of false sites of the same signal, a probability matrix Q is also computed in the same manner. Then, a PWA D representing this type of site is calculated as follows:

$$D^j(x_1 \dots x_k, x_{k+1}) = \log \frac{P^j(x_{k+1}|x_1 \dots x_k)}{Q^j(x_{k+1}|x_1 \dots x_k)} \ . \tag{3.1}$$

PWAs are used to score each potential site along a given sequence. For instance, the score L_D of a potential donor site of length n, $S = s_1 s_2 \dots s_n$ is computed using a first-order ($k = 1$) PWA D as:

$$L_D(S) = \sum_{i=1}^{n-1} D^i(s_i, s_{i+1}) \ . \tag{3.2}$$

This is the log-likelihood ratio of the probability of observing this particular sequence S in a real site versus the probability of observing S in any false site.

3.1.2 Prediction of exons

All potential exons that are compatible with the predicted sites are constructed. By default, only the five highest scoring donor sites that are in frame are considered for each start and acceptor site.

The probability distribution of each nucleotide given the n nucleotides preceding it, is estimated from the exon sequences. The transition probability matrices F^1, F^2 and F^3 are constructed for each one of the three possible reading frames. $F^j(s_1 \dots s_{n+1})$ is the observed probability of finding the sequence $s_1 \dots s_{n+1}$ with s_1 in codon position j. An initial probability matrix I^j is derived from the observed n-mer frequencies at each codon position. From the intron sequences a single transition matrix is computed F_0, as well as a single initial probability matrix I_0. Then, for each $(n+1)$-mer h and frame j the log-likelihood ratio LF is computed as:

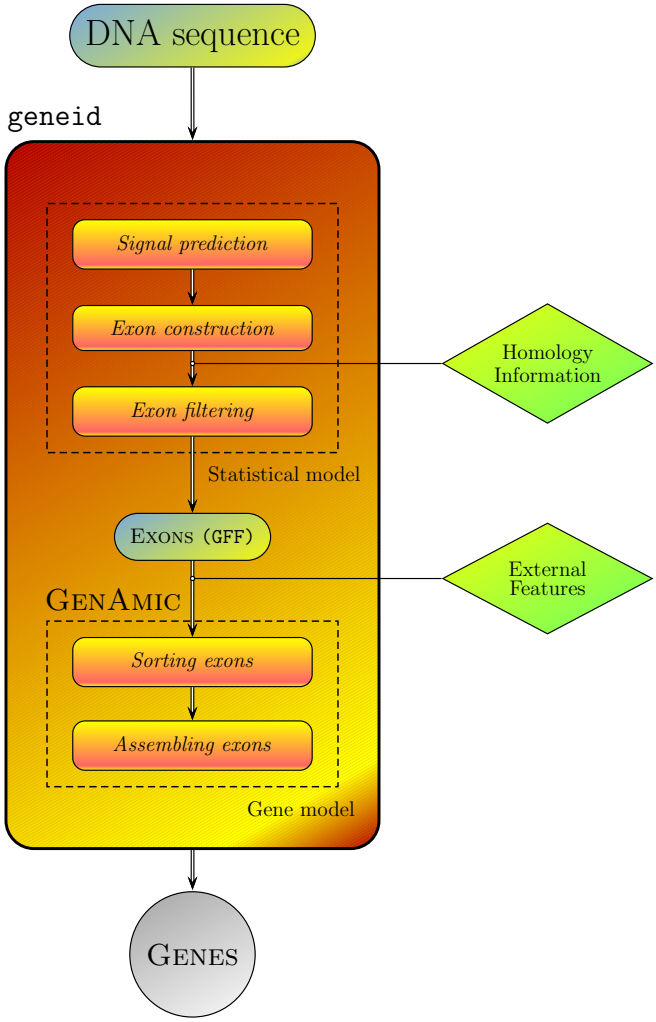


Figure 3.1: General schema of `geneid`: a hierarchical structure that goes from signal recognition and exon building to gene assembly. Adapted from Blanco (2000).

$$LF^j(h) = \log \frac{F^j(h)}{F_0(h)} \ , \tag{3.3}$$

as well as for each n-mer p and frame j the log-likelihood ratio LI is computed as:

$$LI^j(p) = \log \frac{I^j(p)}{I_0(p)} \ . \tag{3.4}$$

Then, given a sequence S of length l in frame j, the protein coding potential L_M of the sequence is defined as:

$$L_M(S) = LI^j(S_1..S_n) + \sum_{i=1}^{l-n} LF^j(S_i..S_{i+n}) \ . \tag{3.5}$$

The final score L_E of a potential exon S, defined by sites s_a (start/acceptor) and s_d (donor/stop) is computed as:

$$L_E(S) = L_A(s_a) + L_D(s_d) + L_M(S) \ . \tag{3.6}$$

This is the log-likelihood ratio of the probability of finding such sites and sequence composition given a real exon over the probability of finding them given a false exon (a real intron).

3.1.3 Gene Model

From a large number of candidate exons, `geneid` selects an appropriate combination of exons to assemble the best gene structure. This assembly must conform to a number of intrinsic biological assumptions such as non-overlap between assembled exons and the maintenance of an open reading frame along assembled genes.

The gene model in `geneid` is the list of rules referring to the succession of elements in the gene structure and to the range allowed distances among them. Each rule is a three column record in the gene model. For instance, the rule

```
First+:Internal+ Internal+:Terminal+ 40:11000
```

indicates that elements (exons) of type *Internal* or *Terminal*, must be immediately assembled after elements of type *First* or *Internal* in the forward strand. The third column indicates the range of valid distances at which these elements can be assembled into a predicted gene. In this case, the elements must be at least 40 bp and at most 11000 bp apart. Users can easily modify the gene model to consider other features such as promoter elements, poly-A tails or secondary structures in the assembly. Such features must then be introduced as external information. The complete `geneid` gene model is shown in Figure 3.2.

```
# GENE MODEL: Rules about gene assembling (GenAmic)
General_Gene_Model
# INTRAgenic connections
First+:Internal+              Internal+:Terminal+           40:11000
Terminal-:Internal-           First-:Internal-              40:11000
# External features
Promoter+                     First+:Single+                50:4000
Terminal+:Single+             aataaa+                       50:4000
First-:Single-                Promoter-                     50:4000
aataaa-                       Single-:Terminal-             50:4000
# INTERgenic conections
aataaa+:Terminal+:Single+     Single+:First+:Promoter+      300:Infinity
aataaa+:Terminal+:Single+     Single-:Terminal-:aataaa-     300:Infinity
Promoter-:First-:Single-      Single+:First+:Promoter+      300:Infinity
Promoter-:First-:Single-      Single-:Terminal-:aataaa-     300:Infinity
```

Figure 3.2: Gene model definition in geneid parameter file.

3.1.4 Assembling genes

geneid constructs genes structures, which can contain multiple genes in both strands. The assembly algorithm tries to optimize the sum of scores of the putative assembled exons. Let g be a gene structure whose sequence of exons is $e_1, e_2, \ldots, e_n$; the scoring function L_G is defined as:

$$L_G(g) = L_E(e_1) + L_E(e_2) + \ldots + L_E(e_n) \ . \tag{3.7}$$

This can be approximately interpreted as the log-likelihood ratio of the probability of the defining sites and the hexamer composition of the resulting product given a gene sequence, over this probability given a non-gene sequence. The gene structure predicted for a given sequence is the gene which maximizes L_G among all gene structures that can be assembled from the set of predicted exons for the sequence.

However, the simple sum of log-likelihoods does not necessarily produce genes with the correct number of exons. If L_E is positive, the genes tend to contain many exons, while if L_E is negative, the genes tend to contain less exons. To overcome this limitation, the score of the exons is corrected by adding a constant EW. Therefore, the new exon scoring function L_E^* is calculated as:

$$L_E^*(S) = L_E(S) + EW \ . \tag{3.8}$$

Given an exon, the parameter EW could be interpreted as the prior odds of being a real exon versus being a false one (Kass and Raftery, 1995). We assume the sequence S, is generated under one of the two hypotheses, being an exon ($exon$) or not being an exon ($\neg exon$). Therefore, given the prior probabilities $p(exon)$ and $p(\neg exon) = 1 - p(exon)$, we are interested in updating our knowledge about how likely this sequence S is an exon in the light of data. This is done by calculating the posterior probabilities $p(exon|S)$ and $p(\neg exon|S)$ and their ratios ,i.e., their posterior odds. By the Bayesian theorem, we can formulate:

$$p(exon|S) = \frac{p(S|exon)p(exon)}{p(S|exon)p(exon) + p(S|\neg exon)p(\neg exon)} \tag{3.9}$$

and

$$p(\neg exon|S) = \frac{p(S|\neg exon)p(\neg exon)}{p(S|exon)p(exon) + p(S|\neg exon)p(\neg exon)} \tag{3.10}$$

so that we obtain the following equation:

$$\frac{p(exon|S)}{p(\neg exon|S)} = \frac{p(S|exon)}{p(S|\neg exon)} \frac{p(exon)}{p(\neg exon)}. \tag{3.11}$$

Then, equation (3.8) would be equivalent to equation (3.11) in log-scale and could be rewritten from a Bayesian perspective as:

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}\ , \tag{3.12}$$

and *EW* could be defined as:

$$EW = log \frac{p(exon)}{p(\neg exon)} \tag{3.13}$$

Thus, if we compute an *EW* with value -7, the ratio of the prior odds would be 1/128. That could be interpreted as the ratio of the probabilities of being an exon versus not being and exon in a exhaustive set of exons generated by `geneid`.

EW must be estimated for each species and for each training set. A simple optimization procedure is performed. Thus, the value that maximizes the accuracy of the predictions in the training set is selected. More formally, the value that maximizes the coefficient of correlation between the actual and the predicted coding nucleotides is selected (as explained in more detail in section 3.3.2).

`geneid` implements the dynamic programming algorithm `genamic` which searches the space of predicted exons in order to assemble the gene structure. From a list of exons, `genamic` computes the best gene ending in every exon and the associated score in a linear time according to the number of input exons (Guigó, 1998).

3.2 Genome Annotation Assessment Project

The Genome Annotation Assessment Project (GASP, Reese et al. (2000), `http://www.fruitfly.org/GASP1/`) was organized by the Berkeley Drosophila Genome Project to formulate guidelines and accuracy standards to evaluate computational annotation tools. The aim of the project was to encourage the development of existing genome annotation approaches through a careful assessment and comparison of the predictions made by all the available programs. The goal of the annotation process is to assign as much information as possible to the raw target sequence with an emphasis on the location of coding genes.

3.2.1 GASP bases

The GASP experiment consisted of the following stages:

- A training data set of curated sequences and the alcohol dehydrogenase (*Adh*) region, including 2.9 Mb of *Drosophila melanogaster* genomic sequence, was collected by the organizers and provided to the participants.
- A set of standard annotations based on experimental data was developed to evaluate submissions while the participating groups produced and submitted their annotations for the region.
- The participant´s predictions were compared to the standards and the results were presented as a tutorial at the Intelligent Systems for Molecular Biology (ISMB, Heidelberg 1999).

The organization chose the 2.9 Mb *Adh* contig because it was large enough to be challenging, contained genes with a variety of sizes and structures, and included regions of high and low gene density.

The annotation used as standard, ideally, should contain the correct structure of all the genes in the region without any error. Unfortunately, such a set was impossible to obtain because the underlying biology of the entire region was incompletely understood. The organization built a two-part approximation to the perfect data set, taking advantage of data from a cDNA sequencing project and a *Drosophila* community effort to build a set of curated annotations for this region (Ashburner et al., 1999). The first standard set, known as *std*1, used high quality sequences from a set of 80 full-length cDNA clones from the *Adh* region to provide a set of annotations that are very likely to be correct but certainly not exhaustive. The second standard set, known as *std*3, was built from the annotations being developed for Ashburner et al. (1999) to give a standard with more coverage of the region, although with less confidence about the accuracy and independence of the annotations.

To evaluate the accuracy of gene prediction in the *Adh* region, *std*1 and *std*3 sets were used. *std*1 is a rigorous annotation set, but incomplete, while *std*3 is as complete as possible, but less reliable. Therefore, the organization decided to compute sensitivity measures using the *std*1 set, and specificity measures to be computed in the *std*3 set. The combination of the two standard sets seemed to sufficiently represent the true nature of the region and conclusions based on them are interesting, and more realistic than previous benchmarks realized on single gene sequences.

The organization also provided several *Drosophila*-specific data sets to enable the participants to tune their tools. The gene curated set, extracted from the Flybase, contained genomic sequences of 275 multi- and 141 single exon non-redundant genes together with their start and stop codons an the splice sites coordinates.

Participants were given the finished sequence for the *Adh* region and the available related training data. However, they did not have access to the full-length cDNA sequences that were sequenced for the paper by Ashburner et al. (1999) that describes the *Adh* region in depth. The experiment was widely announced and open to any participant.

3.2.2 geneid in Drosophila

A special issue of *Genome Research* was dedicated to the GASP, and participants were encouraged to describe their methods and results in detail. Our paper was included in this special issue and describes how the parameters for `geneid` in *D. melanogaster* were computed, the test of different approaches to improve the predictions and the protocol to obtain the final predictions. The final `geneid` predictions showed an accuracy comparable to the gene finding programs that exhibited the highest accuracy in the GASP results published in Reese et al. (2000).

Although `geneid` was not used by the Drosophila Genome Project to annotate the *D. melanogaster* genome, it had some usage through our web page and from people who had freely downloaded the program. Some experimental papers have been based on `geneid` predictions (i.e. Dunlop et al. (2000), Castellano et al. (2001) and Beltran et al. (2003)).

Methods

GeneID in *Drosophila*

Genís Parra, Enrique Blanco, and Roderic Guigó[1]

Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra, E-08003 Barcelona, Spain

GeneID is a program to predict genes in anonymous genomic sequences designed with a hierarchical structure. In the first step, splice sites, and start and stop codons are predicted and scored along the sequence using position weight matrices (PWMs). In the second step, exons are built from the sites. Exons are scored as the sum of the scores of the defining sites, plus the log-likelihood ratio of a Markov model for coding DNA. In the last step, from the set of predicted exons, the gene structure is assembled, maximizing the sum of the scores of the assembled exons. In this paper we describe the obtention of PWMs for sites, and the Markov model of coding DNA in *Drosophila melanogaster*. We also compare other models of coding DNA with the Markov model. Finally, we present and discuss the results obtained when GeneID is used to predict genes in the *Adh* region. These results show that the accuracy of GeneID predictions compares currently with that of other existing tools but that GeneID is likely to be more efficient in terms of speed and memory usage. GeneID is available at http://wwwl.imim.es/~eblanco/GeneId.

GeneID (Guigó et al. 1992) was one of the first programs to predict full exonic structures of vertebrate genes in anonymous DNA sequences. GeneID was designed with a hierarchical structure: First, gene-defining signals (splice sites and start and stop codons) were predicted along the query DNA sequence. Next, potential exons were constructed from these sites, and finally the optimal scoring gene prediction was assembled from the exons. In the original GeneID the scoring function to optimize was rather heuristic: The sequence sites were predicted and scored using position weight matrices (PWMs), a number of coding statistics were computed on the predicted exons, and each exon was scored as a function of the scores of the exon defining sites and of the coding statistics. To estimate the coefficients of this function a neural network was used. An exhaustive search of the space of possible gene assemblies was performed to rank predicted genes according with an score obtained through a complex function of the scores of the assembled exons.

During recent years GeneID had some usage, mostly through a now nonfunctional e-mail server at Boston University (geneid@darwin.bu.edu) and through a WWW server at the IMIM (http://www1.imim.es/geneid.html). During this period, however, there have been substantial developments in the field of computational gene identification (for recent reviews, see Claverie 1997; Burge and Karlin 1998; Haussler 1998), and the original GeneID has become clearly inferior to other existing tools. Therefore, some time ago we began developing an improved version of the GeneID program, which is at least as accurate as other existing tools but much more efficient at handling very large genomic sequences, both in terms of speed and usage of memory. This new version maintains the hierarchical structure (signal to exon to gene) in the original GeneID, but we have simplified the scoring schema and furnished it with a probabilistic meaning: Scores for both exon-defining signals and protein-coding potential are computed as log-likelihood ratios, which for a given predicted exon are summed up into the exon score, in consequence also a log-likelihood ratio. Then, a dynamic programming algorithm (Guigó 1998) is used to search the space of predicted exons to assemble the gene structure (in the general case, multiple genes in both strands) maximizing the sum of the scores of the assembled exons, which can also be assumed to be a log-likelihood ratio. Execution time in this new version of GeneID grows linearly with the size of the input sequence, currently at ~2 Mb per minute in a Pentium III (500 MHz) running linux. The amount of memory required is also proportional to the length of the sequence, ~1 megabyte (MB)/Mb plus a constant amount of ~15 MB, irrespective of the length of the sequence. Thus, GeneID is able to analyze sequences of virtually any length, for instance, chromosome size sequences.

In this paper we describe the "training" of GeneID to predict genes in the genome of *Drosophila melanogaster*. In the context of GeneID training means essentially computing PWMs for splice sites and start codons, and deriving a model of coding DNA, which, in this case, is a Markov model of order 5, similar to the models introduced by Borodovsky and McIninch (1993). Therefore, in the following sections, we describe the training data set used, particularly our attempt to recreate a more realistic scenario to train and test GeneID by generating semiartificial large genomic

[1]**Corresponding author.**
E-MAIL rguigo@imim.es; FAX 34-93-221-3237.

10:511–515 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

contigs from single-gene DNA sequences, and we briefly describe the main features of GeneID for *D. melanogaster*. Then, we present the results obtained in the training data set when different schemas are used to compute scores for sites and coding potential, and the results obtained on the *D. melanogaster Adh* region when the optimal scoring schema in the training set is used to predict genes in this region.

METHODS

Data Sets

We have merged the sets of 275 multi- and 141 single-exon sequences provided by Martin Reese (Reese et al. 2000) as a set of known *D. melanogaster* gene-encoding sequences into the unique MR set. From the MR set we inferred PWMs for splice sites and start codons, and the Markov model of order 5 for coding regions. The MR set contains only single-gene sequences. To assess the accuracy of the predictions in a more realistic scenario, we have randomly embedded the sequences in the MR set in a background of artificial random intergenic DNA as described (R. Guigó, P. Agarwal, J.F. Abril, M. Burset, and J.W. Fickett, in prep.). Thus, a single sequence of 5,689,206 bp embedding the 416 genes in the MR set has been used to evaluate the accuracy of the predictions. The sequence, and the coordinates of the embedded exons are available at http://www1.imim.es/~gparra/GASP1.

GeneID

As outlined, GeneID for *D. melanogaster* uses PWMs to predict potential splice sites and start codons. Potential sites are scored as log-likelihood ratios. From the set of predicted sites (which includes, in addition, all potential stop codons), the set is built of all potential exons. Exons are scored as the sum of the scores of the defining sites, plus the log-likelihood ratio of the Markov model for coding sequences. Finally, the gene structure is assembled from the set of predicted exons, maximizing the sum of the scores of the assembled exons. The procedure is illustrated in Figure 1, which shows the GeneID predictions in a small region of the *Adh* sequence.

Predicting and Scoring Sites

Actual splice sites, and start codons were extracted from the MR set.

Donor Sites

The MR set contains 757 donor sites. From them, a frequency matrix P was derived from position -3 to $+6$ around the exon–intron boundary, with position 0 being the first position in the intron. P_{ij} is the probability of observing nucleotide $i[i \in (A,C,G,T)]$ at position j $[j \in (-3,\ldots,+6)]$ in an actual donor site. The positional frequency Q of nucleotides in the region -3 to $+6$ around all dinucleotides GT was also computed (with position 0 being the position corresponding to the nucleotide G in the GT dinucleotides.) Then, a PWM for donor sites D was calculated as

$$D_{ij} = \log\left(\frac{P_{ij}}{Q_{ij}}\right) \quad (1)$$

PWMs for acceptor sites, A, and start codons, S, were obtained in a similar way. These matrices can be obtained from http://www1.imim.es/~gparra/GASP1.

PWMs can be used to score each potential donor site (GT), acceptor site (AG), and start codon (ATG), along a given sequence. The score of a potential donor site, $S = s_1 s_2 \ldots s_{10}$ within the sequence is computed as

$$L_D(S) = \sum_{i=1}^{10} D_{s_i i} \quad (2)$$

This is the log-likelihood ratio of the probability of observing this particular sequence S in an actual site versus the probability of observing S in any false GT site. Similar scores are computed for acceptor sites (L_A) and start codons (L_B).

Predicting and Scoring Exons

GeneID distinguishes four types of exons: (1) Initial ORFs, defined by a start codon and a donor site; (2) internal ORFs, defined by an acceptor site and a donor site; (3) terminal ORFs, defined by an acceptor site and a stop codon; and (4) single ORFs, defined by a start codon and a stop codon. This corresponds to intronless genes. GeneID constructs all potential exons that are compatible with the predicted sites. (Only the five highest scoring donor sites within frame are considered for each start codon and acceptor site.)

Coding Potential

All exon and intron sequences were extracted from the MR multiexon data set. A Markov model of order 5 was estimated to model both exon and intron sequences, that is, we estimated the probability distribution of each nucleotide given the pentanucleotide preceding it in exon and intron sequences. From the exon sequences we estimated this probability for each of the three possible frames, building the transition probability matrices F^1, F^2, F^3. F^j $(s_1 s_2 s_3 s_4 s_5 s_6)$ is the observed probability of finding hexamer $s_1 s_2 s_3 s_4 s_5 s_6$ with s_1 in codon position j, given that pentamer $s_1 s_2 s_3 s_4 s_5$ is with s_1 in codon position j. An initial probability matrix, I^j, was estimated from the observed pentamer frequencies at each codon position. From the intron sequences a single transition matrix was computed F_0, as well as a single initial probability matrix, I_0. Then, for each hexamer h and frame j a log-likelihood ratio was computed:

$$LF^j(h) = \log \frac{F^j(h)}{F_0(h)} \quad (3)$$

as well as for each pentamer p and frame j

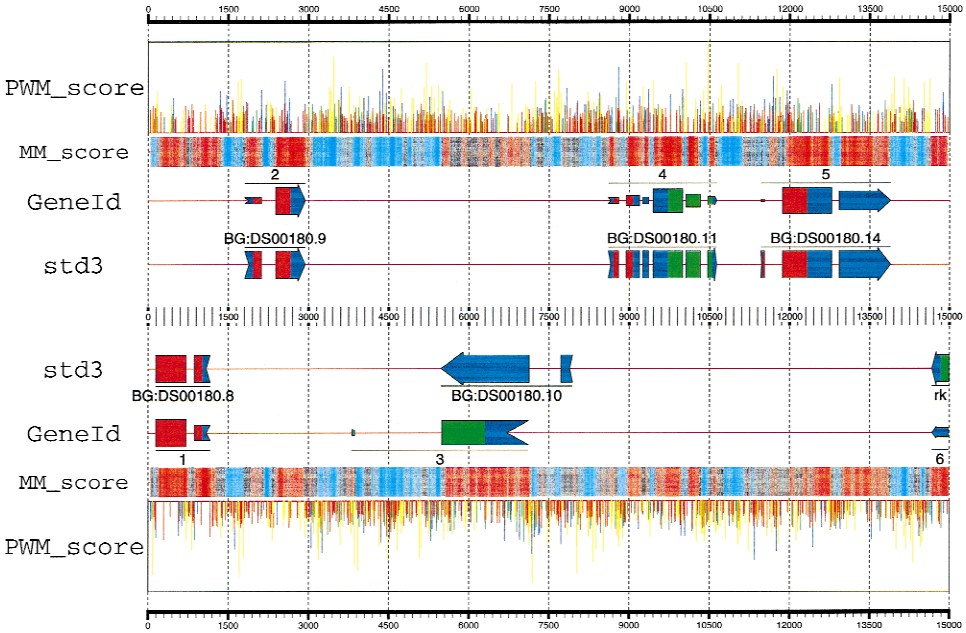


Figure 1 Predictions obtained by GeneID in the region 462500–477500 from the *Adh* sequence, compared with the annotation in the standard std3 set. In a first step, GeneID identifies and scores all possible donor (blue) and acceptor (yellow) sites, start codons (green), and stop codons (red) using PWMs—the height of the corresponding spike is proportional to the site score. A total of 4704 sites were generated along this 15,000-bp region by GeneID, only the highest scoring ones are displayed here. In a second step, GeneID builds all exons compatible with these sites. A total of 11,967 exons were built in this particular region (not displayed). Exons are scored as the sum of the scores of the defining sites, plus the score of their coding potential measured according with a Markov model of order 5. The coding potential is displayed along the DNA sequence (MM_score). Regions strong in red are more likely to be coding than regions strong in blue. From the set of predicted exons, the gene structure is generated, maximizing the sum of the scores of the assembled exons. Exons assembled in the predicted genes are drawn with heights proportional to their scores. A two-color code is used to indicate frame compatibility: Two adjacent exons are frame compatible if the right half of the upstream exon (the remainder) matches the color of the left half of the downstream exon (the frame). Data are from the gff2ps program (available at http://www1.imim.es/~jabril/GFFTOOLS/GFF2PS.html). The input GFF and the configuration files required for gff2ps to generate this diagram can be found at http://www1.imim.es/~gparra/GASP1.

$$LI^j(p) = \log \frac{I^j(h)}{I_0(h)} \tag{4}$$

The distributions F and I can be obtained from http://www1.imim.es/~gparra/GASP1.

Then, given a sequence S of length l in frame j, the coding potential of the sequence is defined as

$$L_M(S) = LI^j(S_{1..5}) + \sum_{i=1}^{l-5} LF^j(S_{i..i+5}) \tag{5}$$

where $S_{i..k}$ is the subsequence of S starting in position i and ending in position k.

The score of a potential exon, S, $L_E(S)$ defined by sites s_a (start/acceptor) and s_d (stop/donor) is computed as

$$L_E(S) = L_A(s_a) + L_D(s_d) + L_M(S) \tag{6}$$

This score can be assumed to be the log-likelihood ratio of the probability of finding such sites and sequence composition given an actual exon over the probability of finding it on a random sequence bounded by AG and GT dinucleotides. Because L_M is the logarithm of the ratio of the probability of the sequence under the coding model over the probability under the noncoding model (not under a random model), L_M only approximates such a log-likelihood ratio.

Assembling Genes

GeneID predicts gene structures, which can be multiple genes in both strands, as sequences of frame-compatible nonoverlapping exons. A minimum intron length of 40 bp and a minimum intergenic distance of 300 bp are enforced. If a gene structure, g, is a sequence of exons, $e_1, e_2, \ldots e_n$, a natural scoring function is

$$L_G(g) = L_E(e_1) + L_E(e_2) + \cdots + L_E(e_n) \quad (7)$$

L_G (g) can be approximately interpreted as the log-likelihood ratio of the probability of the defining sites and the hexamer composition of the resulting product given a gene sequence, over this probability given a nongene sequence. In GeneID, the gene structure predicted for a given sequence is the gene maximizing L_G (g), among all gene structures that can be assembled from the set of predicted exons for the sequence. Because the number of approximations made, the simple sum of log-likelihood ratios does not produce necessarily genes with the correct number of exons (if L_E is positive, the genes tend to have a large number of exons; if L_E is negative, the genes tend to have a small number of exons), and the score of the exons is corrected by adding a constant, IW. Thus, given an exon, e, the actual score of e is

$$L_E^*(e) = L_E(e) + IW \quad (8)$$

To estimate this constant, a simple optimization procedure was performed. Genes were predicted in the training semiartificial genomic sequence for different values of IW, and the value was chosen that maximized the correlation coefficient between the actual and predicted coding nucleotides. This value was found to be $IW = -7$.

RESULTS

Training GeneID

We tested two additional models of coding DNA before deciding for a Markov model of order 5, a Codon usage model, and a model that combined a Markov model of order 1 of the translated amino acid sequence and a Codon preference model (see Guigó 1999 for details on these models). In both cases, log-likelihood ratios were obtained in a similar way to the Markov model log-likelihood ratios (see Methods). For instance, in the case of the Codon usage model, for each triplet s, we estimated the probabilities of the codon s in coding sequences, $U(s)$ and the probability of the triplet in noncoding sequences, $U_0(s)$, and built the log-likelihood ratio

$$LU(s) = \log \frac{U(s)}{U_0(s)}$$

Then, given a sequence, S, of length l in frame 0 (i.e., $S_1S_2S_3$ form a codon), the coding potential of the sequence is computed as

$$L_C(S) = \sum_{i=1,4,7\ldots}^{l-2} LU(S_iS_{i+1}S_{i+2})$$

The models were inferred from the MR set, as the Markov model was, and tested on the MR-set sequences embedded in the large artificial genomic contig. To test the models, genes were predicted using GeneID, but exons were scored using only the scores derived under the coding DNA model (i.e., the scores from the exon defining sites were ignored). Predictions were compared with the annotated genes, and the usual measures of accuracy were computed (Reese et al. 2000). Results are shown in Table 1. For comparison, we also show the results when only the scores of the sites are used to score the exons. As it is possible to see the Markov model of order 5 produces more accurate results than the other models, it was chosen to be used in GeneID to predict the genes in the *Adh* region. As described above, GeneID scores the exons as the sum of the scores of the sites and the Markov model score. Results under this scoring schema, the one effectively used to predict genes in the *Adh* region, are also given in Table 1.

Results in the *Adh* Region

Table 2 shows the results when GeneID, with the parameters estimated above, is used to predict genes in

Table 1. Testing Different Models of Coding DNA in the Training Semiartificial Genomic Sequence

	Base level			Exon level				
	Sn	Sp	CC	Sne	Spe	SnSp	ME	WE
Sites–PWM	0.23	0.65	0.37	0.17	0.13	0.15	0.72	0.79
CU	0.91	0.88	0.88	0.46	0.43	0.45	0.21	0.27
DIA + CP	0.91	0.88	0.89	0.46	0.46	0.46	0.23	0.25
MM-5	0.93	0.90	0.91	0.54	0.51	0.52	0.18	0.24
PWM and MM-5	0.92	0.92	0.92	0.75	0.71	0.73	0.12	0.18

(CU) Codon usage model; (DIA+CP) combination of a Markov model of order 1 of the translated amino acid sequence and a Codon preference model; (MM-5) Markov model of order 5. Genes have been predicted using GeneID, but in each case exons have been scored on the basis solely of the coding DNA model, ignoring the contribution of the exon-defining sites. Predicted genes have been compared with the annotated ones, and the usual measures of accuracy computed. Results obtained when exons are scored as a function only of the scores of the defining sites are also given (Sites–PWM). Finally, we report the results on accuracy when the exons are scored as the sum of the Markov model score and the scores of the exon-defining sites. This is the scoring schema used by GeneID when attempting to predict genes in the *Adh* region.

Table 2. Accuracy of GeneID in the *Adh* Region

	Base level		Exon level					
	Sn (std1)	Sp (std3)	Sn (std1)	Sp (std3)	ME (std1)	WE (std3)	CPU time (sec)	Memory (MB)
GeneID, submitted (1)	0.48	0.84	0.27	0.29	54.4	47.9	74	~500
GeneID, submitted (2)	0.86	0.82	0.59	0.34	21.0	48.0	74	~500
GeneID, current	0.96	0.92	0.70	0.62	11.0	17.0	83	18.11

The std1 annotation data set was used to evaluate sensitivity; the std3 annotation data set to evaluate specificity, as in GASP1 (see Reese et al. 2000). Discrepancies between the accuracy of the submitted predictions, both the initial ones (1) and the corrected (2), and the accuracy of the predictions obtained with the current version of GeneID are due to a number of errors during the process of generating the submitted predictions (see Discussion). The decrease in the amount of memory required to obtain the predictions is due to algorithmic developments occurring after GASP1.

the *Adh* region. Both the results originally submitted to the Genome Annotation Assessment Project (GASP) and the results obtained with the currently available version of GeneID are given (see Discussion). In addition, we provide information on execution time and memory requirements of GeneID to analyze the *Adh* region. The detailed exon coordinates of the predictions by GeneID can be found at http://www1.imim.es/~gparra/GASP1.

DISCUSSION

The results presented above indicate that the current version of GeneID shows an accuracy, as measured by the GASP contest, comparable to the accuracy of the programs based on hidden Markov models (HMMs), which in GASP exhibited the highest accuracy. In favor of GeneID is the simplicity and modularity of its structure, which, as a consequence, is likely to make the program more efficient in terms of speed and memory usage. In GeneID the gene identification problem is stated as a one-dimensional chaining problem for which more efficient algorithms may be designed than for an aligment problem, as gene identification is implicitly formulated in HMMs. Against GeneID is the somehow less rigorous probabilistic treatement of the scoring schema. For instance, we are currently unable to justify the "magic number" (*IW*, see Methods), which needs to be added to the exon scores to obtain accurate predictions.

GeneID submitted rather poor predictions to GASP (see Table 2). Two bugs in the version of the program under development at that time were to blame. They were discovered and a second prediction submitted (see Table 2). After GASP we changed a rather complex schoring schema to the simpler and more natural schema described in Methods, which resulted in higher accuracy. This is the scoring schema currently in use in GeneID.

Although currently fully functional, we are still developing GeneID further. Our short-term plans include, among others, to train GeneID to predict genes in the human and the *Arabidopsis thaliana* genomes and to include the possibility of incorporating the results of database searches—both ESTs and proteins—in the GeneID prediction schema, which can be done rather naturally. The possibility of including external evidence to "force" known genes or exons into the prediction is already included in the working version of GeneID. This may be useful for reannotation of very large genomic sequences. Finally, the current structure of GeneID can be highly parallelized, and we are also working in this direction.

ACKNOWLEDGMENTS

We thank Josep F. Abril and Moisès Burset for helpful discussions and constant encouragement. This work was supported by a grant from Plan Nacional de I+D (BIO98-0443-C02-01) from the Ministerio de Educación y Ciencia (Spain).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Borodovsky, M. and J. McIninch. 1993. Genmark: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17:** 123–113.

Burge, C.B. and S. Karlin. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8:** 346–354.

Claverie, J.M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6:** 1735–1744.

Guigó, R. 1998. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comput. Biol.* **5:** 681–702.

———. 1999. DNA composition, codon usage and exon prediction. In *Nucleic protein databases* (ed. M. Bishop), pp. 53–80. Academic Press, San Diego, CA.

Guigó, R., S. Knudsen, N. Drake, and T.F. Smith. 1992. Prediction of gene structure. *J. Mol. Biol.* **226:** 141–157.

Haussler, D. 1998. Computational genefinding. *Trends in Biochemical Sciences, Supplementary Guide to Bioinformatics: 12–15. Trends Genet.*

Reese, M.G., G. Hartzell, N.L. Harris, U.Ohler, and S.E. Lewis. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* (this issue).

Received February 9, 2000; accepted February 28, 2000.

3.2.3 GASP results

The GASP project gave some insights on the performance of gene prediction programs in large genomic sequences. The accuracy of the different programs is summarized in Table 3.1. Several gene prediction tools had a sensitivity greater than 95% at nucleotide level. There was a great deal of variability at exon level accuracy. Several tools had sensitivity at exon level over 75%. However, their specificity at exon level was generally much lower. The few missing exons combined with the high sensitivity at nucleotide level suggests that several tools were successful at identifying coding regions, but had trouble finding the correct exon boundaries. All the predictors had considerable difficulty with the correct assembly of complete genes. The best tools were able to achieve sensitivities between 0.33 and 0.44. Most programs tend to predict many genes incorrectly. The major problem is the prediction of initial and terminal short coding exons that could be in some cases shorter than 10 bases. Only gene finding tools based on homology searches of databases can predict them.

To summarize the conclusions drawn by the GASP organizers (Reese et al., 2000):

- 95% of the coding nucleotides of the coding genes were correctly predicted.
- The correct structures were predicted for about 40% of the genes. Nucleotide level predictions are easier, exon level predictions are more difficult.
- Current gene prediction programs have achieved major improvements in multiple gene regions.
- Gene finding including ESTs and protein homology does not always improves predictions.
- Programs with specific parameter files for the species under study performed better than the others.
- No program is perfect.

The two last statements encouraged us to continue developing `geneid` and to try to obtain parameter files for the next genomes to be sequenced.

The main conclusions from this experiment were that gene prediction methods had improved and that they could be very useful for whole genome annotations. However, these results could not be extrapolated to, for instance, vertebrates, which have larger genomes and their gene structure is less compact than in *D. melanogaster*. On the other hand, the GASP also showed that high quality annotations depend on a solid understanding of the organism in question.

3.3 Training geneid in other species

As we have seen in the previous section, parameter files derived from a specific species perform better than generic ones. It appears that each genome (each species) has some

	Base level		Exon level			
	Sn (Std1)	Sp (Std3)	Sn (Std1)	Sp (Std3)	WE (Std1)	ME (Std3)
Fgenes CGG1	0.89	0.77	0.65	0.49	10.5	31.6
Fgenes CGG3	0.93	0.60	0.75	0.24	5.6	53.3
GeneMark HMM	0.96	0.86	0.70	0.47	8.1	28.9
Genie	0.96	0.92	0.70	0.57	8.1	17.4
Genie EST	0.97	0.91	0.77	0.55	4.8	20.1
Genie EST HOM	0.97	0.83	0.79	0.52	3.2	22.8
HMM Gene	0.97	0.91	0.68	0.53	4.8	20.2
MAG PIE	0.96	0.63	0.63	0.41	12.1	50.2
Grail exp	0.81	0.86	0.42	0.41	24.3	28.7
geneid	0.96	0.92	0.70	0.61	11.0	17.0

Table 3.1: Evaluation of the different gene finding tools that participated in the GASP. The evaluation is divided into nucleotide level and exon level. From Reese et al. (2000).

characteristic signatures for gene recognition. Under this assumption, we decided to develop training sequence sets and geneid parameter files for species whose genomes were going to be sequenced.

Nowadays, geneid has parameter files for several species. So far, we have developed parameter files for the following species: *Arabidopsis thaliana*, *Ceanorhabditis elegans*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Homo sapiens*(used for other mammal species), *Oryza sativa*, *Plasmodium falciparum*, *Tetraodon nigroviridis* and *Triticum aestivum*. Parameter sets for more species will be available soon. The datasets used to train geneid are freely available at: http://genome.imim.es/datasets/geneid.

3.3.1 Collecting training data

The first step for the development of a parameter file is to gather a set of well annotated sequences: the "training set". The success of the final predictions depends largely on the quality of the data that are used as training set. A good review of the standards to create clean data sets for gene prediction can be found at http://bioinformer.ebi.ac.uk/newsletter/archives/5/gene_prediction.html. A training set is defined as a set of genomic sequences satisfying a number of constraints:

- Genes should have been determined experimentally and not by the outcome of a genome project. The protein should be known or the complete mRNA sequenced.
- For each gene, the genomic sequence should have been sequenced and the coordinates of coding regions exactly mapped.
- The description of the gene must not contain any of the following: alternative gene product, alternative splicing, partial or putative CDS, putative gene, gene prediction nor viral or mithocondrial origin.

- In addition, the sequences must contain the basic structural properties of standard coding genes:
  - Translational start and stop signals should be standard (ATG for start codon and TGA, TAG and TAA for the stop codons).
  - The presence of the minimal canonical signal for the splice sites (with introns starting with GT and ending with AG).
  - The maintenance of the open reading frame through out the translation of the coding exons until the annotated stop codon.

Since the training set is used to derive statistical parameters, the features to model should have a unique representation. To ensure non-redundancy, `blastp` is used to compare all proteins with each other. If any group of protein sequences have a similarity greater than 80% over a strech of 50 amino acids, only one sequence is retained and the others are discarded.

Genomic sequences are mandatory to be able to model splice sites and exonic structure of the genes, however, a set of mRNA sequences is also convenient to complement the amount of coding regions (as described in section 3.3.2).

To gather the training sets we mostly search the EMBL (`http://www.embl.org/`) or GenBank (`http://www.ncbi.nlm.nih.gov/`) databases. For species without enough annotated sequences in the public databases, we contact the consortium in charge of the corresponding genome sequencing project to gather, in collaboration, a reliable set of annotated sequences.

As a result of these collaborations we have participated in the annotation of several genomes. The *Annexed papers* section includes two publications that were partially based on the gene predictions obtained using `geneid` with parameter files specifically developed for each species.

3.3.2 Building the parameter file

Sites definition

To determine which positions are relevant for the definition of a site, the relative entropy is calculated. The positions frequency of nucleotides in the surrounding bases of the canonical signals in both actual exon boundaries P and non functional sites Q is measured (30 base pairs upstream and downstream). Then, for each position j, the relative entropy D_j (also known as the Kullback-Liebler distance (Durbin et al., 1998)) is defined as:

$$D^j(P,Q) = \sum_{i=A,C,G,T} P_{ij}\, log\, \frac{P_{ij}}{Q_{ij}} \quad . \tag{3.14}$$

The stretch of nucleotides crossing the coding exon boundary with the relative entropy above a threshold of 0.1 is taken for the PWA model. After that, the log-likelihood ratio between the real and the non functional site is computed as explained in section 3.1.1.

Depending on the amount of available data for each species and the nature of the signal every type of site could be represented by a PWA of different order. For acceptor sites

in human, a first-order PWA is constructed based on some bias detected in dinucleotides around the canonical signal AG (Burge and Karlin, 1997). However, a second-order PWA is built for start codons to capture the appearance of a second ATG signal after the real one because a biological penalty is known to exist in order to avoid the activation of the second ATG (Kozak, 1999). In contrast, PWAs of order zero, equivalent to PWMs, are constructed for species with less accurate annotations.

Coding potential

First, the sequence of coding regions (CDS) and introns from the training set are extracted. If available, the mRNA set, containing non redundant CDS, can be used in this step to enrich the amount of CDS. Next, the initial and transition matrices for the Markov Model are computed as log-likelihood ratios. The optimal order that reflects the dependencies between contiguous codons seems to be order five. `geneid` allows the use of many orders of Markov chain as a coding statistic. Different orders could be chosen, depending on the amount of available CDS for the corresponding species. In order to create a matrix for order n, $90 * 4^{n+1}$ bases of CDS and $30 * 4^{n+1}$ bases of non-coding sample sequence are required, as estimated by Mark Borodovsky (personal communication). Thus, for a 3rd order matrix you would need at least 23,040 bases of CDS and 7,680 bases of non-coding sequence. Using smaller samples will generate less accurate predictions.

With `geneid`, a complete set of initial and transition probabilities can be incorporated for different C+G content contexts. Thus, signals and exons can be predicted using a different scoring schema, according to their genomic context. For human, three different initial and transition matrices have been constructed depending on the percentage of C+G content (0-45%, 45-55% and 55-100%).

Optimization

A general process of optimization is needed in order to predict the number of real exons. We do not use any maximal optimization algorithm. Rather, we made an extensevely exploration of the *EW* parameter space. A Perl script generates `geneid` predictions from an extesive set of *EW* values (from a minimal and maximal boundaries and with a defined interval). Then, predictions for each *EW* were evaluated and the correlation coefficient is mesured using the actual gene coordinates as reference. The *EW* value that maximizes the correlation coefficient between the actual and the predicted coding nucleotides in the training set is selected.

3.4 Variation in gene structure and splice site signals

The compilation of these training sets is also extremely useful for the comparative analysis of the general mechanisms of gene recognition (including translational, splicing and translational signals). This section tries to describe some structural and compositional properties of gene defining features. Although, nowadays, we have data available for many species, this initial analysis has been done with the paramater files for the following species: *D. melanogaster*, *H. sapiens*, *D. discoideum* and *T. nigroviridis*. Thus, this short

section is a preliminary analysis towards the characterization of peculiarities found in the observed species and further analysis are needed to reach more general conclusions.

Table 3.2 and Figure 3.3 have been generated with the corresponding training sets for each species (described in section 3.2.2 for *Drosophila melanogaster* and in the *Annexed Paper* section for *Dictyostelium discoideum* and *Tetraodon nigroviridis*). The human sequences correspond to the 178 genes used in Guigó et al. (2000).

Unlike the process of mRNA translation by the ribosome, which seems to follow a set of rules that is essentially invariant, the rules governing the RNA splicing clearly differ between different groups of eukaryotes. A graphical representation of the splice sites composition is shown in Figure 3.3. Although all species conserve the canonical GT and AG dinucleotides at the beginning and end of the introns the complete splice signal differs notably. The upstream region of the 3' region (from position -17 to -5), frequently known as poly-pyrimidine track is AT rich in *Dictyostelium*, mostly T rich in *Drosophila* (except for positions -11 to -6 that exhibits a C enrichment), and TC rich in vertebrates. In the region proximal to the 3' exon boundaries, *Dictyostelium* only has conservation in the canonical AG, whereas the other three species seem to have a bias to C in the -3 position and to G in +1 position.

In the donor splice site all species seems to have the conserved motif CAGGTAAGT corresponding to the complementary sequence of the U1 snRNA subunit. However, the overall genomic C+G content seems to model this conservation. In *Dictyostelium*, for instance the positions of the motif containing A and T are more conserved, and even in the first position of the motif that in the other species correspond to A or C, is completely biased to A.

The splice signals in *D. discoideum* show the canonical GT-AG motif. However, in contrast with the other species, besides these common sites only weak preferences for nucleotides adjacent to the donor site could be detected. The distal positions are slightly favored by a (A/T)GT motif. This may be caused by the high mean A/T content in introns of 87%. The splice aparatus has therefore to be able to correctly detect and process this signal in spite of its relative weakness compared to other organisms. Possibly the difference composition content between introns and coding sequences contribute to the recognition of the exon boundaries.

There is considerable variation in the C+G content of exons and introns (Table 3.2). For instance, the average C+G content of the introns in *Dictyostelium* is 9% versus 23% in the entire genome. Therefore, it seems that there is also some constraint on the composition of introns. On the other hand, in vertebrates it seems that exons have a bias in the other direction, coding exons being much richer in C+G than the average in the genome. However, introns have a C+G content more similar to the general genomic composition. This composition bias could play an important role in the identification of introns and exons.

An interesting property observed in the structure of genes (Table 3.2) is the variation of intron and internal coding exon size across the different species. In *Dictyostelium*, intron length seems to have clear restrictions (with a mean of 132 and a standard deviation of 76), whereas the length of the internal coding exons seems to be less restricted (with a average of 544 bp and a standard deviation of 1012). On the other hand, in vertebrates, intron length seems to have no clear restriction (average 641 bp and standard deviation of 975 in human), whereas exons seem to be constrained (average 145 and standard deviation

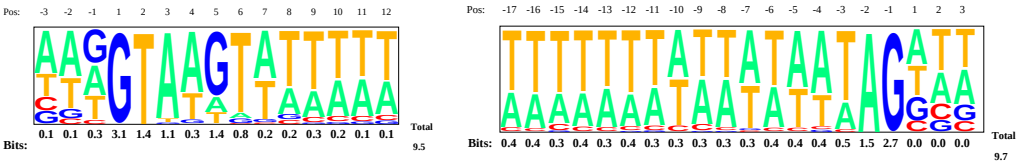
	Exon length (bp)		Intron length (bp)		Exon C+G	Intron C+G	genomic C+G
	average	std. dev.	average	std. dev	(%)	(%)	(%)
D. discoideum	543.92	1012.12	132.02	76.04	0.31	0.09	0.23
D. melanogaster	455.31	618.44	245.92	618.45	0.54	0.38	0.43
T. nigroviridis	140.03	102.65	296.81	670.88	0.53	0.43	0.46
H. sapiens	145.38	95.35	640.85	974.76	0.55	0.45	0.41

Table 3.2: Average exon and intron length and C+G content for the four species under study. Exon refers to internal coding exons

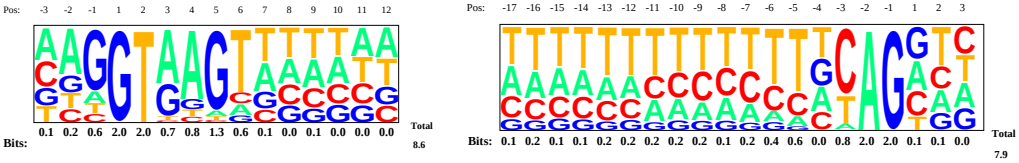
of 95 in human). Intriguingly, *Drosophila* shows intermediate and very variable intron and exon length distributions without any clear pattern of restriction. The most striking observation is that exon length distribution in vertebrates is very similar to the intron length distribution in *Dictyostelium*.

This data is consistent with the differential intron and exon definition where short introns, which are mostly found in lower eukaryotes, seem to be recognized molecularly by the interaction of the splicing factors which bind to both ends of the intron. In vertebrates the internal exons are small (140 nucleotides on average), whereas introns are typically much longer (with some being more than 100 kb). The exon definition was proposed to explain how the splicing machinery recognizes exons in a sea of intronic DNA, where many cryptic splice sites exist. This theory suggests that an internal exon is initially recognized by the presence of a chain of interactions of the splicing factors that bind to it.

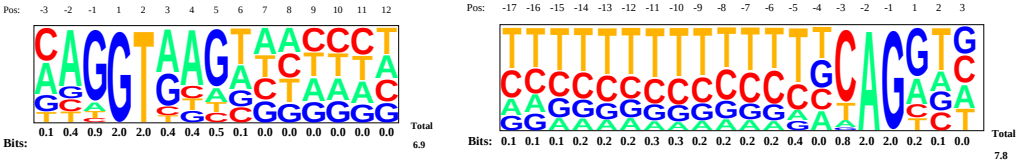
Dictyostelium discoideum



Drosophila melanogaster



Tetraodon nigroviridis



Homo sapiens

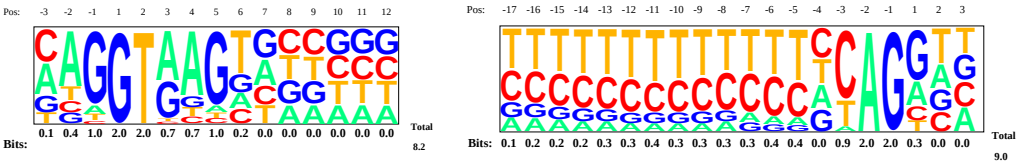


Figure 3.3: Splice signal conservation in different species. Sequence motifs for 5′ splice sites (donors) and 3′ splice sites (acceptors) were generated using Pictogram (`http://genes.mit.edu/pictogram.html`). The height of each letter is proportional to the frequency of the corresponding base at a given position, and bases are listed in descending order from top to bottom. The relative entropy (in bits) of the model relative to the background transcript base composition is also shown.

Comparative gene finding: sgp2

The increasing number of available genomes has lead to the development of new computational gene finding methods that use sequence conservation to improve the accuracy of gene prediction methods (as reviewed in section 1.2.3). Anonymous genomic sequences from different organisms are compared, under the assumption that coding regions tend to be more conserved that non-coding regions. The first part of this section gives a brief overview of the `sgp1` algorithm, its strengths and limitations. Then, we summarize the approaches to overcome these limitations on which `sgp2` is based. This section also includes a short description of `twinscan`. `twinscan` was developed by Korf et al. (2001) and it uses a similar approach to combine comparative information. The attached paper gives a detailed description of `sgp2` structure and the accuracy achieved in different annotated sets of sequences.

4.1 sgp1, Initial Syntenic Gene Prediction

As was mentioned in the *Introduction* (section 1.2.3), there are different ways to exploit the information from genome comparison into gene prediction. The first version of `sgp` (from Syntenic Gene Prediction) was developed mainly by Thomas Wiehe and Roderic Guigó (Wiehe et al., 2001). `sgp1` separates clearly gene prediction from the alignment problem.

`sgp1` starts by aligning two syntenic regions using an external alignment program (for instance `sim90` or `blast`), and then predicts the final gene structures in which the exons are compatible with the alignment. A central strategy of `sgp1` is to rely as little as possible on species specific nucleotide composition, such as isochore distribution, codon bias or any other coding statistic. Therefore, predicted exons do not receive scores that depend on any of such sources of information. Rather, scoring at the initial step (before the alignment) relies exclusively on splice site quality.

After the alignment, `sgp1` generates a set of pairs of pre-candidate exons between the two species. A pre-candidate exon is a sequence with a well defined reading frame and splice signals. A filtering process checks whether the begin and end positions of any pair of pre-candidates are contained in the alignment regions. If there is any discrepancy, the pair is discarded. Optionally, the filter can be relaxed to allow for an offset between alignment and pre-candidate exon. There are two parameters: x, the number of base pairs

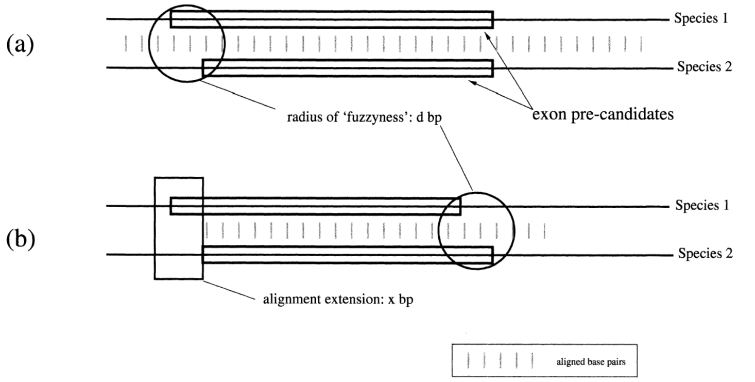


Figure 4.1: Relaxed filtering of pre-candidate exons in `sgp1`.(a) Non exact exon boundaries, but complete coverage by the alignment. (b) Non exact exon boundaries and partial coverage by the alignment. Setting parameters d and x to a value greater than 0 retains pre-candidates with unaligned splice sites. Adapted from Wiehe et al. (2001)

by which locally aligned segments are extended, and d, the maximal distance by which the ends of two paired pre-candidates may be separated (see Figure 4.1).

The exons that pass the filter are assembled into gene predictions independently for both species using the chaining algorithm described by Guigó (1998). The assembly program attempts to build complete gene models consisting of either a single exon or one initial exon, an arbitrary number of internal exons, and one terminal exon. Multiple genes, on either strand, can be assembled.

Given two sequences and their alignment as input, the program calls subroutines for the alignment post-processing, generating pre-candidate exons, exon filtering, rescoring and gene assembly and for generating the final output. The subroutine with the highest time complexity is the one that filters the pre-candidate exons. A very rough estimation of its running time is $O(nm)$, quadratic time, where n and m are the lengths of the input query sequences. This is due to the fact that the size of the two exon pre-candidates lists depends on n and m, respectively, and each pair of pre-candidates, one from each list, has to be processed. This is one of the major limitations of `sgp1` for whole genome predictions. The amount of comparisons of exons structures increases quadratically with the length of the sequences. Therefore, this is a computationally very expensive approach to comparative prediction in complete eukaryotic genomes.

Another important limitation of `sgp1` is that it relies too much on syntenic sequences. If any of the sequences is partially sequenced the accuracy of the method drops substantially. This limits, again, its utility when analyzing complete, large, eukaryotic genomes. In particular when one genome is in non-assembled shotgun form.

4.2 New strategies to overcome sgp1

To overcome these limitations, sgp2 takes a different approach. Essentially, the query sequence from the target genome is compared against a collection of sequences from the informant genome (which can be a single sequence homologous to the query sequence, a whole assembled genome, or a collection of shotgun reads). The results of the comparison are used to modify the scores of the exons produced by geneid *ab initio* gene prediction program.

One of the most important differences between sgp1 and sgp2, is that sgp2 does not attempt to generate all the compatible exons of the two orthologous sequences. Finding compatible exons requires that genes in the two sequences have the same exon-intron structure. Extending this strategy to multi-gene sequences would require the assumption that the two sequences have the same genes in the same order and orientation. In a large-scale comparison there are a lot of partial duplications and rearrangements, and even sequencing mis-assemblies that complicate such approaches.

Using a global alignment or the compatible exons strategy requires informant sequences to be finished. The sequence conservation approach taken by sgp2, which is based on the highest scoring local alignments, allows to use draft and shotgun sequences. The sequence conservation effectively rearranges the alignments into the correct order and orientation. In addition because local alignments can be from any region of the informant genome, it allows us to take, apart from the similarities from orthologies, the similarity observed from paralogies or domain conservation occurring in non syntenic regions.

sgp2 combines tblastx genome comparison results with geneid. tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. To score the alignment an amino acid substitution matrix is used. An amino acid substitution matrix is a 20 x 20 matrix in which every possible identity and substitution is assigned a score based on the observed frequencies of such occurrences in alignments of related proteins. Scores are computed as log-likelihood ratios. Identities are assigned the most positive scores. Frequently observed substitutions also receive positive scores and less observed substitutions are given negative scores (see Figure 4.2). Therefore, tblastx is much more sensitive that using blastn alone. tblastx can capture similarities at amino acid level that could be difficult or impossible to find at nucleotide level. Moreover, the amino acid substitution matrices score the alignments from an evolutionary point of view.

After the tblastx alignment, the maximum scores of the overlapping high-scoring segment pairs (HSPs) are projected in the target sequence in the *maximum scoring projection*. The *maximum scoring projection* is provided to geneid in general feature format (GFF, http://www.sanger.ac.uk/Software/formats/GFF/) where each line contains the coordinates of the alignment and the highest observed score. Essentially, geneid is used to predict all potential exons along the target sequence. Scores of exons are computed as log-likelihood ratios, function of the splice sites defining the exon, the coding bias in composition of the exon sequence as measured by a Markov Model of order five, and of the optimal alignment at the amino acid level between the target exon sequence and the counterpart homologous sequence in the reference set. From the set of predicted exons, the gene structure is assembled (potentially multiple genes in both

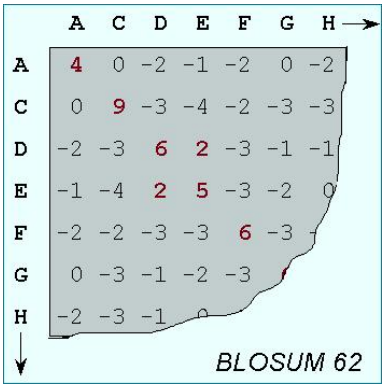


Figure 4.2: A section of the 20 x 20 BLOSUM62 matrix in which every possible identity and substitution is assigned a score based on the observed frequencies of such occurrences in alignments of related proteins.

strands), maximizing the sum of the scores of the assembled exons.

A similar approach has also been recently explored by Korf et al. (2001) for their program `twinscan`. In `twinscan`, the genome sequences are compared using `blastn` and the results serve to modify the underlying probability of the potential exons predicted by `genscan`. `genscan` assigns one of the possible sequence states to each nucleotide of an input sequence (see Figure 1.8). In `twinscan`, `genscan` model that assigns a probability to any parsed DNA sequence is combined with a parallel sequence conservation model. Coding, UTR, and intron/intergenic states are assigned probability to stretches of sequence conservation using a 5th order Markov Model. Models of sequence conservation at splice donor and acceptor sites were based on a 2nd order PWA. These models are not based on dependencies between nucleotides but on dependencies in the pattern of conservation.

To summarize, `twinscan` takes as input local alignments between a target genome and a database of sequences from an informant genome. For each nucleotide of the target genome, only the highest scoring HSP overlapping that nucleotide is used. These alignments are converted into a representation called *conservation sequence*, which assigns one of the three symbols to each possible nucleotide in the alignment: "|" if the alignment contains a match, ":" if it is a gap or a mismatch , and "." if there is no overlapping alignment (as shown in Figure 4.3). Given a target genomic sequence and the *conservation sequence* model, `twinscan` predicts the more probable gene structures according to the probability of corresponding to a particular state together with the given pattern of conservation.

4.3 sgp2: Comparative gene prediction in human and mouse

The following paper gives a more detailed description of the `sgp2` algorithm, specially to the *maximum scoring projection* of the HSPs obtained with the `tblastx` and the re-

Figure 4.3: Conversion of the best local alignment in each region of the target genome (top) into the *conservation sequence* representation used by `twinscan` (bottom). A typical coding region (left), in which there are no unaligned bases or gaps, and the distance between mismatches tend to be multiple of three. A typical intron (right), in which there are unaligned regions, gaps and adjacent mismatches. Adapted from Korf et al. (2001).

scoring of the `geneid` exons. It also shows the evaluation of `sgp2` in different single genes sets of sequences and in the human chromosome 22. Finally, whole genome human and mouse `sgp2` predictions are analyzed.

Methods

Comparative Gene Prediction in Human and Mouse

Genís Parra,[1] Pankaj Agarwal,[2] Josep F. Abril,[1] Thomas Wiehe,[3] James W. Fickett,[4] and Roderic Guigó[1,5]

[1]Grup de Recerca en Informàtica Biomèdica. Institut Municipal d'Investigació Medica / Universitat Pompeu Fabra / Centre de Regulació Genòmica 08003 Barcelona, Catalonia, Spain; [2]GlaxoSmithKline, King of Prussia, Pennsylvania 19406, USA; [3]Freie Universität Berlin and Berlin Center for Genome Based Bioinformatics (BCB), 14195 Berlin, Germany; [4]AstraZeneca R&D Boston, Waltham, Massachusetts 02451, USA

The completion of the sequencing of the mouse genome promises to help predict human genes with greater accuracy. While current ab initio gene prediction programs are remarkably sensitive (i.e., they predict at least a fragment of most genes), their specificity is often low, predicting a large number of false-positive genes in the human genome. Sequence conservation at the protein level with the mouse genome can help eliminate some of those false positives. Here we describe SGP2, a gene prediction program that combines ab initio gene prediction with TBLASTX searches between two genome sequences to provide both sensitive and specific gene predictions. The accuracy of SGP2 when used to predict genes by comparing the human and mouse genomes is assessed on a number of data sets, including single-gene data sets, the highly curated human chromosome 22 predictions, and entire genome predictions from ENSEMBL. Results indicate that SGP2 outperforms purely ab initio gene prediction methods. Results also indicate that SGP2 works about as well with 3x shotgun data as it does with fully assembled genomes. SGP2 provides a high enough specificity that its predictions can be experimentally verified at a reasonable cost. SGP2 was used to generate a complete set of gene predictions on both the human and mouse by comparing the genomes of these two species. Our results suggest that another few thousand human and mouse genes currently not in ENSEMBL are worth verifying experimentally.

After the genome sequence of an organism has been obtained, the very first next step is to compile a complete and accurate catalog of the genes encoded in this sequence. For higher eukaryotic organisms, however, the accuracy of currently available gene prediction methods to perform such a task is limited (Guigó et al. 2000; Rogic et al. 2001; Guigó and Wiehe 2003). The increasing availability of genome sequences from different organisms, however, has lead to the development of new computational gene finding methods that use sequence conservation to help identifying coding exons, and improve the accuracy of the predictions (Fig. 1; Crollius et al. 2000; Wiehe et al. 2000; Miller 2001; Rinner and Morgenstern 2002). Indeed, three such comparative gene prediction programs, SLAM (Pachter et al. 2002), SGP2, and TWINSCAN (Korf et al. 2001) have been used for the comparative analysis of the human and mouse genomes. These analyses lead to more accurate gene predictions, and to the verification of previously unconfirmed genes. In this paper, we describe the program SGP2. Typical computational ab initio gene prediction methods rely on the identification of suitable splicing sites, start and stop codons along the query sequence, and the computation of some measure of coding likelihood to predict and score candidate exons, and delineate gene structures (see Claverie 1997; Burge and Karlin 1998; Haussler 1998; Zhang 2002 and references therein for reviews on computational gene finding).

Similarity between the query sequence and known coding sequences (amino acid or cDNA) can also be used to infer gene structures. When the query sequence encodes a protein for which a close homolog exists, a special type of alignment can be used between the DNA sequence and the target protein/cDNA sequence, in which gaps in the target sequence corresponding to introns in the query sequence must be compatible with potential splicing signals. This is the approach in GENEWISE (Birney and Durbin 1997) and PROCRUSTES (Gelfand et al. 1996). Alternatively, the results of searching the query sequence against a database of known coding sequences, using for instance BLASTX (Altschul et al. 1990, 1997; Gish and States 1993), can be incorporated more or less ad hoc into the scoring schema of an ab initio gene prediction method. The program GENOMESCAN (Yeh et al. 2001), which incorporates BLASTX search results into the predictions by the GENSCAN program (Burge and Karlin 1997), is an example of a recent development in that direction.

Recently developed comparative gene prediction programs further exploit sequence similarity. Instead of comparing anonymous genomic sequences to known coding sequences, anonymous genomic sequences are compared to anonymous genomic sequences from the same or different organisms, under the assumption that regions conserved in the sequence will tend to correspond to coding exons from homologous genes. The approach taken by the different programs to exploit this idea differs notably.

In one such approach (Blayo et al. 2002; Pedersen and Scharl 2002), the problem is stated as a generalization of pairwise sequence alignment: Given two genomic sequences coding for homologous genes, the goal is to obtain the predicted exonic structure in each sequence maximizing the score of the

[5]Corresponding author.
E-MAIL rguigo@imim.es; FAX 34 93 224-0875.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.871403.

13:108–117 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

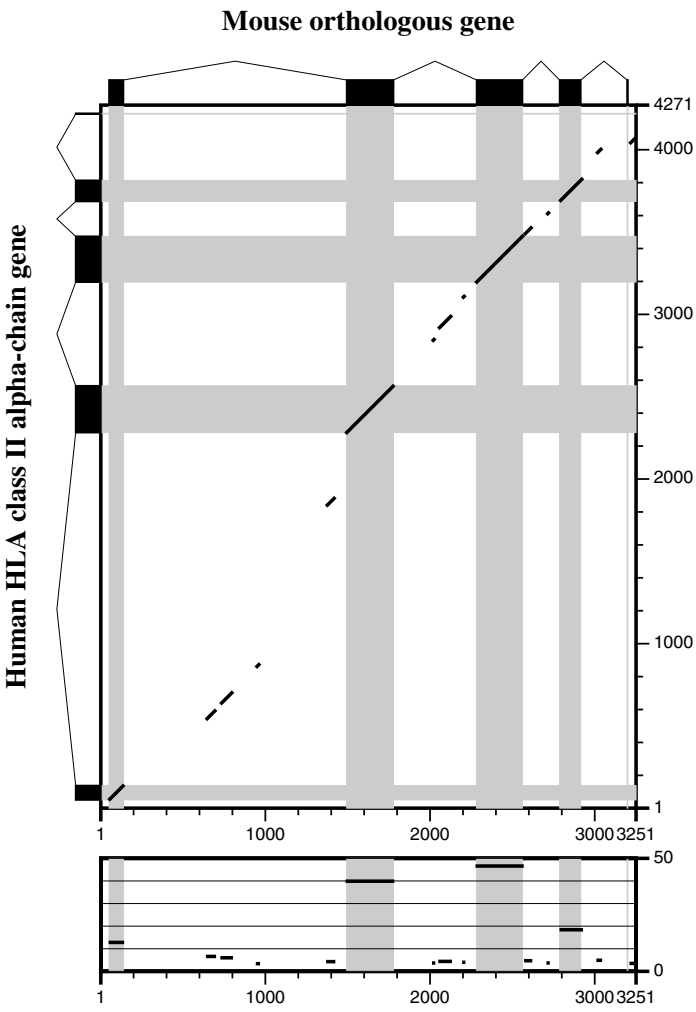


Figure 1 Pairwise comparison using TBLASTX of the human and mouse genomic sequences coding for the HLA class II alpha chain. Black boxes indicate the coding exons, while black diagonals indicate the conserved alignments. The score of the conserved alignments (divided by 10) is given in the lower panels. Although conserved regions between the human and mouse genomic sequences coding for these genes fully include the coding exons, a substantial fraction of intronic regions is also conserved. The TBLASTX outptut was post-processed to show a continuous non-overlapping alignment.

alignment of the resulting amino acid sequences. Both Blayo et al. (2002) and Pedersen and Scharl (2002) solve the problem through a complex extension of the classical dynamic programming algorithm for sequence alignment.

In a different approach, the programs SLAM (Pachter et al. 2002) and DOUBLESCAN (Meyer and Durbin 2002) combine sequence alignment pair hidden Markov Models (HMMs; Durbin et al. 1998) with gene prediction generalized HMMs (GHMMs; Burge and Karlin 1997) into the so-called generalized pair HMMs. In these, gene prediction is not the result of the sequence alignment, as in the programs above; gene prediction and sequence alignment are obtained simultaneously.

A third class of programs adopt a more heuristic approach, and separate clearly gene prediction from sequence alignment. The programs ROSSETA (Batzoglou et al. 2000), SGP1 (from 'syntenic gene prediction'; Wiehe et al. 2001), and CEM (from 'conserved exon method'; Bafna and Huson 2000) are representative of this approach. All these programs start by aligning two syntenic sequences and then predict gene structures in which the exons are compatible with the alignment. The programs described thus far rely on the comparison of fully assembled (and when from different organisms, syntenic) genomic regions. This limits their utility when analyzing complete large eukaryotic genomes, and in particular when the informant genome is in nonassembled shotgun form. To overcome this limitation, the programs TWINSCAN (Korf et al. 2001) and SGP2 take still a different approach. The approach is reminiscent of that used in GENOMESCAN (Yeh et al. 2001) to incorporate similarity to known proteins to modify the GENSCAN scoring schema. Essentially, the query sequence from the target genome is compared against a collection of sequences from the informant genome (which can be a single homologous sequence to the query sequence, a whole assembled genome, or a collection of shotgun reads), and the results of the comparison are used to modify the scores of the exons produced by ab initio gene prediction programs. In TWINSCAN, the genome sequences are compared using BLASTN, and the results serve to modify the underlying probability of the potential exons predicted by GENSCAN. In SGP2, the genome sequences are compared using TBLASTX (W. Gish, 1996–2002, http://blast.wustl.edu), and the results are used to modify the scores of the potential scores predicted by GENEID. TWINSCAN and SGP2 have been successfully applied to the annotation of the mouse genome

(Mouse Genome Sequencing Consortium 2002), and have helped to identify previously unconfirmed genes (Guigó et al. 2003).

In the next section, we describe the algorithmic details of SGP2, and its implementation. We also describe the sequence sets used to benchmark SGP2 accuracy. Results based on these data sets indicate that SGP2 is an improvement over pure ab initio gene prediction programs, even when the informant genome is only in shotgun form. We have found that 3x coverage will generally suffice to achieve maximum accuracy. Finally, we describe the application of SGP2 to the comparative analysis of the human and mouse genomes.

METHODS

SGP2

SGP2 is a method to predict genes in a *target* genome sequence using the sequence of a second *informant* or *reference* genome. Essentially, SGP2 is a framework to integrate the ab initio gene prediction program GENEID (Guigó et al. 1992; Parra et al. 2000) with the sequence similarity search program TBLASTX. The approach is conceptually similar to that used in TWINSCAN to incorporate BLASTN searches into GENSCAN.

GENEID is a genefinder that predicts and scores all potential coding exons along a query sequence. Scores of exons are computed as log-likelihood ratios, which are a function of the splice sites defining the exon, and of the coding bias in composition of the exon sequence as measured by a Markov Model of order five (Borodovsky and McIninch 1993). From the set of predicted exons, GENEID assembles the gene structure (eventually multiple genes in both strands), maximizing the sum of the scores of the assembled exons, using a dynamic programming chaining algorithm (Guigó 1998).

When using an informant genome sequence to predict genes in a target genome sequence, ideally we would like to incorporate into the scores of the candidate exons predicted along the target sequence, the score of the optimal alignment at the amino acid level between the target exon sequence and the counterpart homologous exon in the informant genome sequence. If a substitution matrix, for instance from the BLOSUM family, is used to score the alignment, the resulting score can also be assumed to be a log-likelihood ratio: informally, the ratio between the likelihood of the alignment when the amino acid sequences code for functionally related proteins, and the likelihood of the alignment, otherwise. In principle, this score could be added to the GENEID score for the exon. TBLASTX provides an appropriate shortcut to often find a good enough approximation to such an optimal alignment, and infer the corresponding score: The optimal alignment can be assumed to correspond to the maximal scoring high-scoring segment pairs (HSP) overlapping the exon. However, when dealing in particular with the informant genome sequence in fragmentary shotgun form, often different regions of a candidate exon sequence will align optimally to different informant genome sequences. Thus, in the approach used here, we identify the optimal HSPs covering each fraction of the exon, and compute separately the contribution of each HSP into the score of the exon. In the next section, we describe in detail how this computation is performed.

Scoring of Candidate Exons

Let e be one of the candidate exons predicted by GENEID along the query DNA sequence S. In SGP2, the final score of e, $s(e)$, is computed as

$$s(e) = s_g(e) + ws_t(e)$$

where $s_g(e)$ is the score given by GENEID to the exon e, and $s_t(e)$ is the score derived from the HSPs found by a TBLASTX search overlapping the exon e. Both scores are log-likelihood ratios (and we compute both base two). Assuming that both components are independent, they can be summed up into a single score. However, the assumption of independence is not realistic, $s_g(e)$ depends on the probability of the sequence of e, assuming that e codes for a protein, while $s_t(e)$ depends on the probability of the optimal alignment of e with a sequence fragment of the mouse genome, assuming that both sequences code for related proteins. Obviously, these two probabilities are not independent. Their joint distribution could only be investigated—at least empirically—if the Markov Model of coding DNA used in GENEID, and the substitution matrix used by TBLASTX were inferred from the very same set of coding sequences. Since this is quite difficult, if not unfeasible, we use an "ad hoc" coefficient, w, to weight the contribution of TBLASTX search, $s_t(e)$ into the final exon score.

We compute $s_t(e)$ in the following way. Let $h_1 \cdots h_q$ be the set of HSPs found by TBLASTX after comparing the query sequence S against a database of DNA sequences (Fig. 2A).

First, we find the *maximum scoring projection* of the HSPs onto the query sequence. We simply register the maximum score among the scores of all HSPs covering each position, and then partition the query sequence in equally maximally scoring segments (bounded by dotted lines in Fig. 2A) $x_1 \cdots x_r$, with scores $s_p(x_1) \cdots s_p(x_r)$ (Fig. 2B).

Then, for each predicted exon e (Fig. 2C), we find X_e, the set of maximally scoring segments overlapping e

$$X_e = \{x_i : x_i \cap e \neq \emptyset\}$$

where $a \cap b$ denotes the overlap between sequence segments a and b, and $\emptyset$ means no overlap. We compute $s_t(e)$in the following way:

$$s_t(e) = \sum_{x \in X_e} s_p(x) \frac{|x \cap e|}{|x|}$$

where $|a|$ denotes the length of sequence segment a.

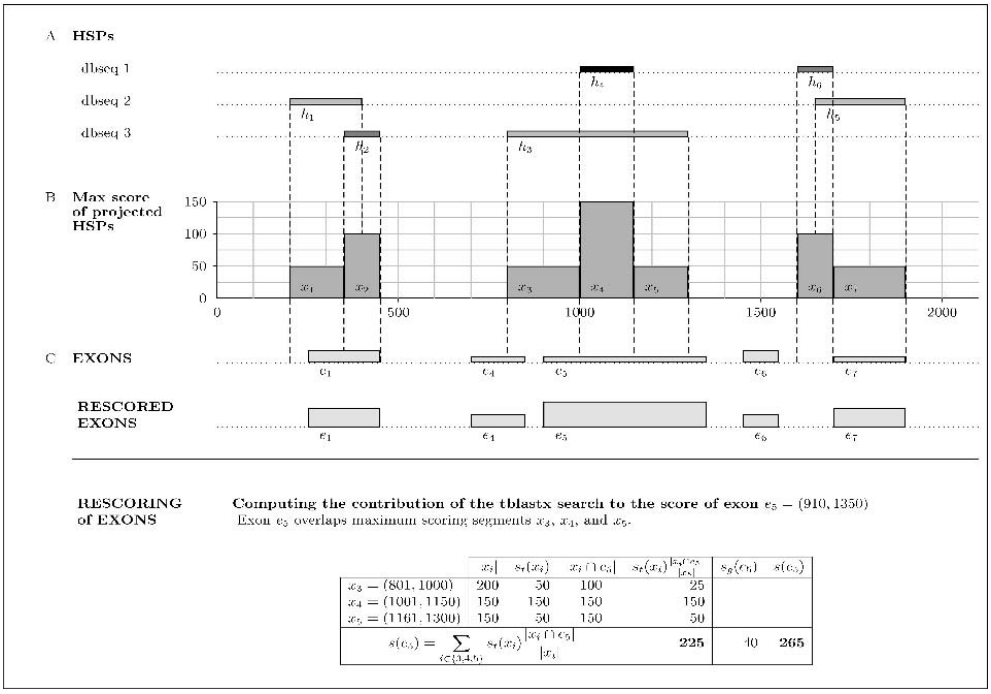
That is, each exon gets the score of the maximally scoring HSPs along the exon sequence proportional to the fraction of the HSP covering the exon. In other words, $s_t(e)$ is the integral of the maximum scoring projection function within the exon interval.

Once the scores s have been computed for all predicted exons in the sequence S, gene prediction proceeds as usual in GENEID: The gene structure is assembled maximizing the sum of scores of the assembled exons.

Running SGP2

In practice, we run SGP2 in the following way. Given a DNA query sequence and a collection of DNA sequences, we compare the query sequence against the collection using TBLASTX 2.0MP-WashU [23-Sep-2001]. The query sequence can be a genomic fragment of any size, including complete eukaryotic chromosomes, whereas the collection of sequences may be almost anything from just a homologous region or a partial collection of genomic sequences from the same or another species to the whole genome sequence of a second species, either completely assembled or in shotgun form at any degree of coverage. In particular, two different regions of the same genome coding for homologous genes can be used within SGP2; in this case the same genome acts as target and informant.

In all the analyses reported here, we used BLOSUM62 as the amino acid substitution matrix, but changed the penalty for aligning any residue to a stop codon to -500. This helps to get rid of a large fraction of HSPs in noncoding regions. Because of TBLASTX limitations, large query sequences may need to be split in fragments before the search, and the results reconstructed afterwards. Results of TBLASTX search are then



	$\lvert x_i \rvert$	$s_t(x_i)$	$\lvert x_i \cap e_5 \rvert$	$s_t(x_i)^{\frac{\lvert x_i \cap e_5 \rvert}{\lvert x_i \rvert}}$	$s_g(e_5)$	$s(e_5)$
$x_3 = (801, 1000)$	200	50	100	25		
$x_4 = (1001, 1150)$	150	150	150	150		
$x_5 = (1161, 1300)$	150	50	150	50		
$s(e_5) = \sum_{i \in \{3,4,5\}} s_t(x_i)^{\frac{\lvert x_i \cap e_5 \rvert}{\lvert x_i \rvert}}$				**225**	40	**265**

Figure 2 Rescoring of the exons predicted by GENEID according to the results of a TBLASTX search. See the "SGP2" section for a detailed explanation of the figure.

parsed to obtain the *maximum scoring projection* of the HSPs onto the query sequence. The parsing includes discarding all HSPs below a given bit score cutoff, subtracting this value from the score of the remaining HSPs, weighting the resulting score by *w* (see above), and collapsing the HSPs in to the maximum scoring projections. In all analyses described here, the bit score cutoff was set to 50, and *w* to 0.20. These values were chosen to optimize the gene predictions in sequence sets of known homologous human and mouse genomic sequences (see the Results section).

The *maximum scoring projection* is given to GENEID in general feature format (GFF; R. Durbin and D. Haussler, http://www.sanger.ac.uk/Software/GFF/). GENEID uses it to rescore the exons predicted along the query sequence as explained, and assembles the corresponding optimal gene structure. GENEID was already designed to incorporate external information into the gene predictions, and no changes were required in the program to accommodate it into the SGP2 context, only a small adjustment in the parameter file to cope with the change in scale of the exon scores.

We have written a simple PERL script which, given a query DNA sequence and the results of the TBLASTX search, performs all the components of the SGP2 analysis transparently: the parsing of the TBLASTX search results, and the GENEID predictions. In the case wherein both the query and the informant sequence are single genomic fragments, the gene predictions can be obtained in both sequences (without the need for a second TBLASTX search). The script, as well as the individual components, can be found at http://www1.imim.es/software/sgp2/.

GENEID has essentially no limits to the length of the input sequence, and deals well with chromosome size sequences. Limits to the length of the input query sequence that can be analyzed by SGP2 are, thus, those imposed by TBLASTX. GENEID is quite fast; given the parsed TBLASTX results, it takes 6 h to reannotate the whole human genome in a MOSIX cluster containing four PCs (PentiumIII Dual 500 Mhz processors).

Accelerating TBLASTX Searches

TBLASTX searches, although efficient, are much slower. Its default usage may become computationally prohibitive when comparing complete eukaryotic genomes. In the context of SGP2, however, a number of TBLASTX options can be changed to speed up the search, without significant loss of sensitivity in the predictions (see the Results section). Thus, results in human chromosome 22 and whole-genome comparisons have been performed using the following set of parameters: W = 5, -nogap, -hspmax = 150,000, B = 200, V = 200, E = 0.01, E2 = 0.01, Z = 30,000,000, -filter = xnu + seg, and S2 = 80. In these cases, the query sequences have been broken up in 5 MB fragments, and the database sequences in 10 MB fragments. In all cases, stop codons are heavily penalized (-500) in the alignments. After the search is completed, locations of the resulting HSPs are recomputed in chromosomal coordinates. Results in the single-gene sequence benchmark data sets were obtained with default TBLASTX parameters.

Sequence Data Sets

Benchmark Sequence Sets

To optimize some of the parameters in SGP2 and to test its performance, we used a set of known pairs of genomic sequences coding for homologous human and rodent genes. The set is built after the set constructed by Jareborg et al. (1999). This is a set of 77 orthologous mouse and human gene pairs. We considered only the 33 pairs of sequences in this set

coding for single complete genes. In addition, we discarded six additional pairs, when we suspected that one of the members could be wrongly annotated. Orthology in the Jareborg et al. (1999) data set is based on sequence conservation. This could bias the set towards the more highly conserved human/mouse orthologous genes. To compensate for this bias, we obtained an additional set of pairs of human/rodent orthologous genes through an approach which does not involve sequence conservation: We obtained the set of pairs of human/mouse sequences from the SWISSPROT database sharing the prefix (indicating the gene) in their locus names. We kept only those pairs for which it was possible to find the corresponding annotated genomic sequence—including the mapping of the transcript, and not only of the coding regions—in the EMBL database. Fifteen additional genes were found this way. Three of them were discarded because we suspected wrong annotation in at least one of the members of the pair. We believe that orthology in the remaining cases is highly likely because of the absolute conservation of the exonic structure (number and length of exons, and intron phases) that we observed. We will call the resulting concatenated set of 39 pairs of human/mouse homologous genes the SCIMOG dataset (from Sanger Center IMim Orthologous Genes). The data set and the detailed protocol used to obtain it can be accessed at http://www1.imim.es/datasets/sgp2002/.

To test the accuracy of SGP2, we used the data set constructed by Batzoglou et al. (2000) of 117 orthologous human and mouse genes. We discarded those pairs in which in at least one of the sequences contained multiple genes, and those in which the coding region started in position 1 in one of the sequences of the pair. This resulted in 110 genes. We will call this set the MIT data set. There is some overlap between the SCIMOG and MIT data sets, and thus the latter cannot properly be called a test set. However, we decided not to eliminate the redundant entries, so that the results could be compared to those published for the ROSSETA program (Batzoglou et al. 2000).

Finally, we tested SGP2 in the complete sequence of human chromosome 22 (Dunham et al. 1999). The masked sequence was obtained from http://genome.cse.ucsc.edu/goldenPath/22dec2001/. Chromosome 22 is probably the best annotated human chromosome. We used the gene annotations at http://www.cs.columbia.edu/~vic/sanger2gbd/. The CDS set contains 554 genes. This is a conservative set that only contains the coding region of genes and does not include pseudogenes. This may lead to an underestimation of the specificity of the predictions.

Mouse and Human Genome Sequences

We used versions MGSCv3 of the mouse genome (2,726,995,854 bp, http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/) and NCBI28 of the human genome (3,220,912,202 bp, http://genome.cse.ucsc.edu/goldenPath/22dec2001/). Both masked and unmasked sequences were obtained from these locations. ENSEMBL gene annotations for these genomes were obtained from http://genome.cse.ucsc.edu/goldenPath/22dec2001/database/ensGene.txt.gz for the human genome, and from http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/database/ensGene.txt.gz for the mouse genome. ENSEMBL predicts 23,005 and 22,076 nonoverlapping transcripts genes on the human and mouse genome, respectively.

Evaluating Accuracy

The measures of accuracy used here are extensively discussed in Burset and Guigó (1996). We will restate them briefly. Accuracy is measured at three different levels: nucleotide, exon, and gene. At the nucleotide and exon levels, we compute essentially the proportion of actual coding nucleotides/exons that have been correctly predicted—which we call *sensitivity*—and the proportion of predicted coding nucleotides/exons that are actually coding nucleotides/exons—which we call *specificity*. To compute these measures at the exon level, we will assume that an exon has been correctly predicted only when both its boundaries have been correctly predicted. To summarize both *sensitivity* and *specificity*, we compute the *correlation coefficient* at the nucleotide level, and the average of *sensitivity* and *specificity* at the exon level. At the exon level, we also compute the *missing exons*, the proportion of actual exons that overlap no predicted exon, and the *wrong exons*, the proportion of predicted exons that overlap no real exons.

At the gene level, a gene is correctly predicted if all of the coding exons are identified, every intron–exon boundary is correct, and all of the exons are included in the proper gene. In addition, we compute the missed genes (MGs), real genes for which none of its exons are overlapped by a predicted gene, and the wrong genes (WGs), predictions for which none of the exons are overlapped by a real gene. In general, gene finders predict the initial and terminal exons very poorly. This often leads to so-called chimeric predictions—one predicted gene encompassing more than one real gene—or to split predictions—one real gene split in multiple predicted genes. Reese et al. (2000) developed two measures, split genes (SG) and joined genes (JG), to account for these tendencies. SG is the total number of predicted genes overlapping real genes divided by the number of genes that were split. Similarly, JG is the total number of real genes that overlap predicted genes divided by the number of predicted genes that were joined.

RESULTS

Benchmarking SGP2

We evaluated the accuracy of SGP2 using a number of different data sets. The lack of a gold standard of gene prediction makes it difficult to get accurate assessments from any single data set. We primarily used three data sets as described earlier.

To benchmark SGP2, we constructed BLAST databases from the mouse and human sections of SCIMOG and MIT, and each mouse/human sequence to the entire human/mouse database, respectively. This enabled us to predict genes in both the mouse and human databases. The results from

Table 1. Gene Prediction in the SCIMOG Data Set

	Nucleotide			Exon				
Program	Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE
GENSCAN	0.98	0.86	0.92	0.84	0.75	0.79	0.04	0.14
TBLASTX default	0.89	0.76	0.81	0.81	—	—	0.19	0.11
SGP2 (single complete genes)	0.97	0.98	0.97	0.89	0.89	0.89	0.03	0.03
SGP2 (multiple genes)	0.94	0.97	0.95	0.80	0.87	0.83	0.10	0.02

Table 2. Gene Prediction Accuracy in the MIT Data Set

	Nucleotide			Exon				
Program	Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE
GENSCAN	0.98	0.89	0.93	0.82	0.75	0.78	0.06	0.13
ROSSETA	0.95	0.97	—	—	—	—	0.02	0.03
TBLASTX default	0.94	0.79	0.85	—	—	—	0.13	0.13
SGP2 (single complete genes)	0.97	0.98	0.97	0.84	0.85	0.84	0.05	0.03
SGP2 (multiple genes)	0.96	0.97	0.96	0.71	0.79	0.75	0.12	0.03

comparing SGP2, GENSCAN, and ROSSETA accuracy values in this case are taken from Batzoglou et al. (2000), and the results of a simple TBLASTX search on the MIT data set are in Table 2 (below). For the TBLASTX searches, the *maximum scoring projection* of the HSPs (see the above section titled "SGP2") was assumed to be the gene prediction. The score cutoff for the HSPs was chosen to maximize the correlation coefficient (CC) between the projected HSPs and the coding exons. In Table 1,2, we report the accuracy of GENSCAN, SGP2, and TBLASTX on the SCIMOG dataset. The accuracy values for SGP2 are reported under two scenarios: assuming a single complete gene and assuming multiple genes. Both GENEID and SGP2 allow the external specification of a *gene model* (i.e., a small number of rules specifying the legal assemblies of exons into gene structures). These rules can be used to force SGP2 to predict a single complete gene to make the results comparable to those of ROSSETA. Without such a restriction (i.e., making no assumptions about the number and completeness of the genes potentially encoded in the query sequence), the results are more directly comparable to those of GENSCAN (although GENSCAN also has a tendency to start a prediction in any sequence with an initial exon, and to terminate it with a terminal exon).

The accuracy of SGP2 is comparable to that of ROSSETA, and is significantly higher than that of GENSCAN. SGP2 also improves substantially over a simple TBLASTX search. The relative low specificity of the TBLASTX search—even after the large penalties for stop codons—reflects the fact that a substantial fraction of the conservation between the human and mouse genomes extends into the noncoding regions (Mouse Genome Sequencing Consortium 2002). At the nucleotide level, SGP2 accuracy is almost equal in the MIT data set and the SCIMOG data set (even though the SGP2 was trained on SCIMOG). The accuracy at the exact exon level, however, decreases, in particular when prediction of multiple genes is allowed. This is a problem inherited from GENEID, which tends to replace short initial and terminal exons with longer internal exons.

Accuracy of SGP2 as a Function of the Coverage of the Mouse Genome

To investigate the utility of partial shotgun data as informant sequence in our approach based on TBLASTX, we simulated shotgun mouse sequence data at different levels of coverage (1.5x, 3x, and 6x) from the mouse genes in the SCIMOG data set, and used them to compare the human sequences in SCIMOG using TBLASTX. The mouse genomic sequences was shredded with uniformly distributed length between 500 and 600 bp with random starting points. No sequencing errors were introduced. At each coverage, we measured the CC between the TBLASTX hits projected along the human genome sequence, and the coding exons (choosing the TBLASTX score cutoff resulting in the optimal CC). With 1.5x coverage, a substantial fraction of the human coding region is not identified by TBLASTX, whereas with 3x, the results are quite similar to those obtained with 6x, which are identical to those obtained with the fully assembled syntenic regions (Table 3). This indicates that even with 3x coverage of the informant genome, our method will produce results nearly identical to those obtained with fully assembled regions. Assembled genomes, however, result in faster TBLASTX searches.

Accuracy of SGP2 in Human Chromosome 22

Human chromosome 22 was the first human chromosome fully sequenced (Dunham et al. 1999), and it is quite the best annotated thus far, due to a number of experimental followups (Das et al. 2001; Shoemaker et al. 2001). Therefore, it provides an excellent data set to validate any gene prediction technology. Human chromosome 22 was searched using TBLASTX against the masked whole-genome assembly from the mouse genome (MGSCv3). The HSPs in chromosomal coordinates resulting from the TBLASTX search were used in GENEID to perform SGP2 gene prediction. Although the HSPs had been computed on the masked sequence, in this case the SGP2 predictions were obtained on the unmasked one. SGP2 predicted 729 genes on human chromosome 22. Table 4 shows the comparative accuracy of the SGP2, GENSCAN, GENOMESCAN, and pure ab initio GENEID predictions (without TBLASTX data). GENSCAN predictions on the masked sequence were taken from the USCS genome browser http://genome.cse.ucsc.edu/. GENOMESCAN predictions were obtained from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/build28_chr_genomescan.gtf.gz. Pure ab initio GENEID predictions were obtained on the masked sequence, and can also be downloaded from http://www1.imim.es/genepredictions/.

Although SGP2 is not more sensitive than GENSCAN, it appears to be more specific (as it utilizes the mouse genome).

Table 3. Accuracy of TBLASTX Predictions as a Function of the Degree of Coverage in the SCIMOG Data Set

	Nucleotide			Exon	
Coverage	Sn	Sp	CC	ME	WE
Simulated 1.5x	0.79	0.78	0.77	0.25	0.10
Simulated 3x	0.86	0.76	0.80	0.21	0.11
Simulated 6x	0.89	0.76	0.81	0.19	0.11
Fully assembled	0.89	0.76	0.81	0.19	0.11

Table 4. Accuracy of Gene-finding Programs on Human Chromosome 22

	Nucleotide			Exon					Gene						
Program	Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE	Sn	Sp	(Sn+Sp)/2	MG	WG	JG	SG
GENSCAN	0.86	0.50	0.64	0.70	0.40	0.55	0.13	0.50	0.06	0.04	0.05	0.11	0.45	1.24	1.07
GENOMESCAN	0.87	0.44	0.59	0.72	0.36	0.54	0.10	0.55	0.11	0.06	0.08	0.12	0.52	1.07	1.14
GENEID	0.80	0.63	0.69	0.66	0.53	0.59	0.19	0.35	0.09	0.07	0.08	0.14	0.39	1.20	1.08
TBLASTX	0.84	0.39	0.54	—	—	—	0.12	0.74	—	—	—	0.11	—	—	—
SGP2	0.83	0.67	0.73	0.68	0.56	0.62	0.16	0.31	0.13	0.10	0.11	0.14	0.36	1.14	1.13

Fifty percent of the GENSCAN-predicted exons do not overlap annotated chromosome 22 exons; this number is only 31% for SGP2. Overall, SGP2 appears to be more accurate than GENSCAN in human chromosome 22: GENSCAN's CC at the nucleotide level is 0.64, whereas that of SGP2 is 0.73. Although accuracy decreases for both programs when going from single-gene sequences (Tables 1, 2) to an entire chromosome, SGP2 retains more accuracy. GENSCAN overall shows higher sensitivity than SGP2, but there were 45 real genes not predicted by GENSCAN on human chromosome 22, and SGP2 was able to predict, at least partially, 15 of them. This suggests that SGP2 and GENSCAN may play complementary roles. GENOMESCAN, on the other hand, did not appear to be superior to GENSCAN in human chromosome 22.

Mouse matches (TBLASTX HSPs) covered 11% of the human chromosome 22. Though they covered 85% of the coding nucleotides, 74% of the HSPs fell outside annotated coding regions. This illustrates the difficulties of using genome sequence conservation even at the protein level between human and mouse genomes to infer coding genes.

Prediction of Genes in the Human and Mouse Genomes

We used SGP2 to predict the entire complement of human (NCBI28) and mouse (MGSCv3) genes. The masked sequences of these two genomes were compared using TBLASTX. The TBLASTX HSPs were used within SGP2. SGP2 predicted 44,242 genes in the human genome, and 44,777 genes in the mouse genome. Obviously, it is difficult to accurately assess these predictions. We used ENSEMBL genes as the set of reference annotations and compared both GENSCAN and SGP2 predictions to it. Figure 3 shows summaries of the accuracy of SGP2 at the chromosome level in the human and mouse genomes. When compared against ENSEMBL, SGP2 is more accurate than GENSCAN.GENSCAN. It is more specific at the nucleotide level: the average SGP2 specificity is 0.60 for human and 0.61 for mouse, whereas these values for GENSCAN are 0.43 and 0.44. SGP2 is also equally sensitive at the nucleotide level: The average SGP2 sensitivity is 0.82 for human and 0.85 for mouse; these values for GENSCAN are 0.82 and 0.84. Overall, the average SGP2 CCs are 0.70 for human and 0.72 for mouse, and for GENSCAN, the respective averages are 0.59 and 0.61. The accuracy of the SGP2 predictions, moreover, appears to be more consistent across chromosomes than that of the GENSCAN predictions. Interestingly, human chromosome Y is an outlier, with genes in this chromosome being poorly predicted. Genes in chromosome Y appear to be more difficult to predict than genes in other chromosomes for pure ab initio gene prediction programs, because chromosome Y is also an outlier for GENSCAN. SGP2 suffers, in addition, on human chromosome Y because the mouse chromosome Y has yet to be sequenced, and thus there was no comparative information available.

Overall, 23,913 of the human predictions and 24,203 of the mouse predictions overlapped ENSEMBL genes, whereas 95% of the mouse and 93% of the human ENSEMBL genes were among the genes predicted by SGP2. Of the remaining putative novel 20,570 mouse SGP2 genes and 20,193 human SGP2 genes, 10,456 mouse and 9,006 human predictions were found to be similar at $P < 10^{-6}$ to a prediction in the counterpart genome. Of these, 5,960 and 4,909 have multiple exons and are longer than 300 bp. A significant fraction of these putative homologous predictions are likely to correspond to real genes (Guigó et al. 2003). The predictions are interactively accessible through the USCS genome browser (http://genome.cse.ucsc.edu/) and through the DAS server at ENSEMBL (http://www.ensembl.org, under "DAS sources"). The complete set of prediction files is available at http://www1.imim.es/genepredictions/.

Speeding Up TBLASTX Searches

Using TBLASTX to compare human and mouse whole-genome sequences, even in masked form, is quite expensive computationally because of the 6-frame translation on both query and target. To substantially reduce the search time, we used a word size of 5 and sacrificed some sensitivity (see the section above titled "Accelerating TBLASTX Searches" for details). We also penalized stop codons heavily and did not permit gaps. The computation took an estimated 500 CPU days on a farm of Compaq Alphas.

Accuracy in Tables 1 and 2 was computed using default TBLASTX parameters. Table 5 shows the comparative accuracy of TBLASTX and SGP2 predictions, under the default and the speed-up configuration of TBLASTX parameters on the SCIMOG data set. The sensitivity of speed-up TBLASTX searches drops from 0.89 to 0.72, but specificity increases slightly. SGP2 is more robust, and it compensates for some of the sensitivity lost in the TBLASTX search. Overall accuracy for SGP2, as measured by the CC, drops only from 0.95 to 0.93.

Predictions on human chromosome 22 and the whole human and mouse genomes have been obtained with this speed-up configuration of parameters.

DISCUSSION

We have described the program SGP2 for comparative gene finding, and presented the results of its application to the human and mouse genome sequences. Results in controlled benchmark sequence data sets indicate that, by including in-

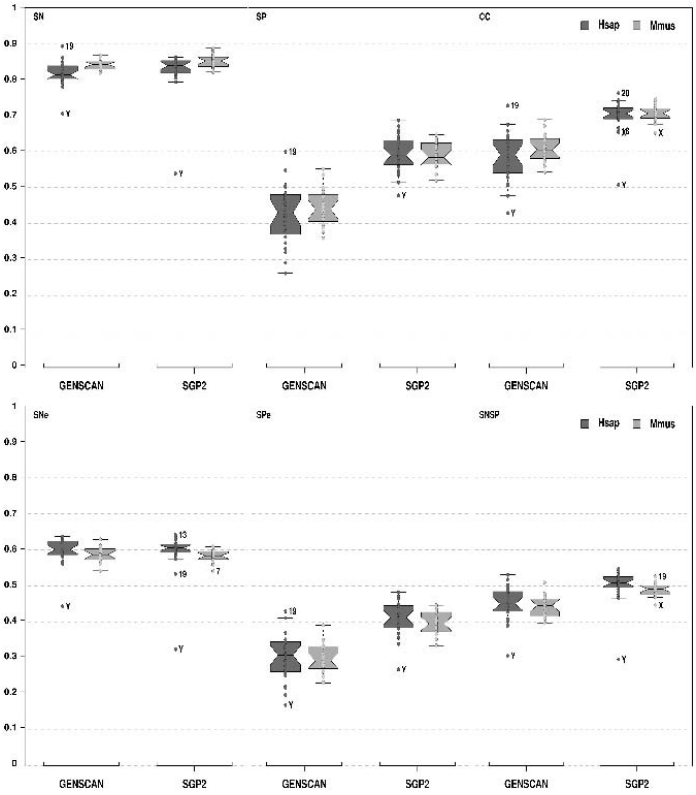


Figure 3 Accuracy of the human and mouse SGP2 and GENSCAN predictions. The accuracy was measured in the entire chromosome sequences using the standard accuracy measures: SN, (sensitivity); SP, (specificity); CC, (correlation coefficient); SNe, (exon sensitivity); SPe, (exon specificity); and SNSP, (average of sensitivity and specificity at exon level). Predictions from both programs were compared against the human and mouse ENSEMBL annotations. Each dot corresponds to the accuracy measure of one chromosome. Chromosome labels are shown for outlier values. The boxplots (Tukey 1977) were obtained using the R-package (http://cran.r-project.org/).

formation from genome sequence conservation, predictions by SGP2 appear to be more accurate than those obtained by pure ab initio programs, exemplified here by GENSCAN and GENEID. Although there is not a significant gain in sensitivity, the specificity of the predictions appears to increase substantially, and a smaller number of false positive exons are predicted.

Indeed, one the major obstacles towards the completion of the catalog of human (mammalian) genes is our inability to assess the reliability of the large number of computational gene predictions that have not been verified experimentally. Whereas the ENSEMBL pipeline produces about 25,000 human and mouse genes, the NCBI annotation pipeline predicts almost 50,000 genes in mouse, and the program GENOMESCAN predicts close to 55,000 genes in this species. Although a large fraction of the ENSEMBL genes correspond to computational predictions without experimental verification, the method is quite conservative, and recent experiments suggest that essentially all ENSEMBL genes are indeed real (Guigó et al. 2003). The problem remains with the tens of thousands of additional computational predictions that are not included in ENSEMBL. A fraction of them are likely to be real, but the question is how large this fraction is. The results obtained here in human chromosome 22 seem to indicate that it may not be very large. Although the existence of hundreds of unidentified genes in this chromosome cannot be completely ruled out, the results strongly suggest that a substantial fraction of these additional computational gene predictions are false positives.

In this regard, the results presented here demonstrate that through the comparison of the human and mouse genomes using SGP2 (or another available comparative gene prediction tool), the false-positive rate can be reduced significantly, and the catalog of mammalian genes better defined. SGP2 predicts a few thousand candidate genes not in ENSEMBL that we believe are worth verifying experimentally. Indeed, the experimental verification of a subset of these provides evidence of at least 1000 previously nonconfirmed genes (Guigó et al. 2003).

The predictions by SGP2 obtained here are, of course, still far from definitively setting this catalog. For one thing, the mouse may be too close a species to human: A large fraction of the sequence has been conserved between the genomes of these two species. Indeed, most sequence conservation between human and mouse does not correspond to coding exons (Mouse Genome Sequencing Consortium 2002), compounding gene prediction. This suggests that the genome of another vertebrate species evolutionarily located between fish and mammals could be of great utility towards closing in the vertebrate (and mammalian) gene catalog.

SGP2 is flexible enough so that it can be easily accommodated to analyze species other than human and mouse. The fact that it can deal with shotgun data at any level of coverage means that as the sequence of a new genome starts becoming available, it can be used to improve the annotation of other already existing genomes. Particularly relevant in this context is a feature of SGP2 (and GENEID) that we have not explored here. SGP2 can produce predictions on top of preexisting annotations. For instance, we could have given to SGP2 the location and exonic coordinates (in GFF format) of known REFSEQ genes (or ENSEMBL), and SGP2 would have predicted genes only outside the boundaries of these genes of

Table 5. Accuracy of TBLASTX and SGP2 Predictions Using "Default" versus Speed-Up Parameters

		Nucleotide			Exon				
		Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE
Default	TBLASTX	0.89	0.76	0.81	—	—	—	0.19	0.11
	SGP2	0.94	0.97	0.95	0.80	0.87	0.83	0.10	0.02
Speed-up	TBLASTX	0.72	0.80	0.75	—	—	—	0.22	0.10
	SGP2	0.88	0.98	0.93	0.77	0.85	0.81	0.12	0.02

already well known exonic structure. Preliminary results indicate that this approach improves gene prediction outside of the preassumed genes, and reduces the rate of chimeric predictions (i.e., predictions encompassing multiple genes). Moreover, we believe that SGP2 can be substantially improved. The flexibility of the SGP2/GENEID framework makes it quite easy to integrate additional information that can contribute to the accuracy of the predictions: synonymous versus nonsynonymous substitution rates in the alignments by TBLASTX, conservation of the splice signals in the informant genome, amino acid substitution matrices specific to the phylogenetic distance between the species compared, etc.

In this regard, the reasons to use the default BLOSUM62 matrix are not obvious. Given the expected sequence similarity between mouse–human orthologs, BLOSUM80 appears to be a better choice. However, we intended to also detect divergent families. Towards that end, the superiority of BLOSUM80 is less clear. We have compared TBLASTX search results on human chromosome 22 against the whole mouse genome. Whereas the HSPs resulting from the BLOSUM62 search cover 84% of the chromosome 22 coding nucleotides, BLOSUM80 HSPs cover 88% of them. However, BLOSUM80 is much less specific than BLOSUM62: 60% of the nucleotides in the BLOSUM62 HSPs fall outside coding regions, compared to 88% for BLOSUM80. It is thus clear that the optimal matrix or combination of matrices for comparative gene-finding using TBLASTX requires further investigation.

Although a large fraction of the human genome sequence has been known for more than a year, the exact number of human genes and their precise definition remain unknown. Gene specification in higher eukaryotic sequences is the result of the complex interplay of sequence signals encoded in the primary DNA sequence, which is only partially understood. Without an exhaustive catalog of human genes, however, the promises of genome research in medicine and technology cannot be completely fulfilled. The work presented here, in which it is shown that human–mouse comparisons can contribute to the completion of the mammalian (human) gene catalog, underscores the importance of the comparisons of the genomes of different organisms to fully understand the phenomenon of life, and in particular to deciphering the mechanism, central to life, by means of which the genome DNA sequence specifies the amino acid sequence of the proteins.

ACKNOWLEDGMENTS

We thank the Mouse Genome Sequencing Consortium for providing the mouse genome sequence as well as support throughout the analysis process. We especially thank Francisco Câmara for arranging the data listed in the gene-prediction page on our group Web site, and for setting up and taking care of our DAS server. We also thank Ian Korf for inspiring discussions regarding the parameters to use in the TBLASTX search. We thank Enrique Blanco, Sergi Castellano, and Moisés Burset for helpful discussions and constant encouragement. This work was supported by a grant from Plan Nacional de I+D (BIO2000-1358-C02-02), Ministerio de Ciencia y Tecnologia (Spain), and from a fellowship to J.F.A. from the Instituto de Salud Carlos III (99/9345).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bafna, V. and Huson, D.H. 2000. The conserved exon method. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8:** 3–12.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.

Birney, E. and Durbin, R. 1997. Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 56–64.

Blayo, P., Rouzé, P., and Sagot, M.-F. 2002. Orphan gene finding—An exon assembly approach. *Theoretical Computer Science* (in press).

Borodovsky, M. and McIninch, J. 1993. GenMark: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17:** 123–134.

Burge, C.B. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Burge, C.B. and Karlin, S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8:** 346–354.

Burset, M. and Guigó, R. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353–357.

Claverie, J.-M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6:** 1735–1744.

Crollius, H.R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* **25:** 235–238.

Das, M., Burge, C.B., Park, E., Colinas, J., and Pelletier, J. 2001. Assessment of the total number of human transcription units. *Genomics* **77:** 71–78.

Dunham, I., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Durbin, R., Eddy, S., Crogh, A., and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of protein and nucleic acids*. Cambridge University Press, Cambridge.

Gelfand, M.S., Mironov, A.A., and Pevzner, P.A. 1996. Gene recognition via spliced alignment. *Proc. Natl. Acad. Sci.* **93:** 9061–9066.

Gish, W. and States, D. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3:** 266–272.

Guigó, R. 1998. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comp. Biol.* **5:** 681–702.
Guigó, R. and Wiehe, T. 2003. Gene prediction accuracy in large DNA sequences. In *Frontiers in computational genomic* (eds. M.Y. Galperin and E.V. Koonin), Caister Academic Press, Norfolk, UK.
Guigó, R., Knudsen, S., Drake, N., and Smith, T.F. 1992. Prediction of gene structure. *J. Mol. Biol.* **226:** 141–157.
Guigó, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. 2000. Gene prediction accuracy in large DNA sequences. *Genome Res.* **10:** 1631–1642.
Guigó, R., Dermitzakis, E.T., Agarwal, P., Pontig, C.P., Parra, G., Reymond, A., Abril, J.F., Keibler, E., Lyle, R., Ucla, C., et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci.* (in press).
Haussler, D. 1998. Computational genefinding. *Trends in biochemical sciences, supplementary guide to bioinformatics*, pages 12–15.
Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.
Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1:** 140–148.
Meyer, I.M. and Durbin, R. 2002. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* **18:** 1309–1318.
Miller, W. 2001. Comparison of genomic DNA sequences: Solved and unsolved problems. *Bioinformatics* **17:** 391–397.
Mouse Genome Sequencing Consortium 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Pachter, L., Alexandersson, M., and Cawley, S. 2002. Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comp. Biol.* **9:** 389–400.
Parra, G., Blanco, E., and Guigó, R. 2000. Geneid in *Drosophila*. *Genome Res.* **10:** 511–515.
Pedersen, C. and Scharl, T. 2002. Comparative methods for gene structure prediction in homologous sequences. In *Algorithms in Bioinformatics* (eds. R. Guigó, and D. Gusfield), Springer-Verlag, Berlin, Germany.
Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10:** 483–501.
Rinner, O. and Morgenstern, B. 2002. Agenda: Gene prediction by comparative sequence analysis. *In Silico Biol.* **2:** 0018.
Rogic, S., Mackworth, A.K., and Ouellette, F. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11:** 817–832.
Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* **409:** 922–927.
Tukey, J.W. 1977. *Exploratory data analysis*. pp. 39–41. Addison-Wesley, Boston, MA.
Wiehe, T., Guigó, R., and Miller, W. 2000. Genome sequence comparisons: Hurdles in the fast lane to functional genomics. *Brief. Bioinform.* **1:** 381–388.
Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T., and Guigó, R. 2001. SGP-1: Prediction and validation of homologous genes based on sequence alignments. *Genome Res.* **11:** 1574–1583.
Yeh, R., Lim, L., and Burge, C. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11:** 803–816.
Zhang, M.Q. 2002. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* **3:** 698–709.

WEB SITE REFERENCES

http://www.sanger.ac.uk/Software/formats/GFF/; GFF format description page.
http://genome.cse.ucsc.edu/goldenPath/22dec2001/; Human genome sequence goldenpath from Dec. 22, 2001 (hg10) equivalent to NCBI28 build.
http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/; Mouse genome sequence goldenpath from Feb. 2002 (mm2) equivalent to MGSCv3.
http://www.cs.columbia.edu/~vic/sanger2gbd; Victoria Haghighi, Human chromosome 22 curated annotations.
ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/build28_chr_genomescan.gtf.gz; Genomescan predictions from NCBI.
http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/database/ensGene.txt.gz; Mouse ENSEMBL annotations file.
http://blast.wustl.edu; Washington University BLAST Archives
http://genome.cse.ucsc.edu/goldenPath/22dec2001/database/ensGene.txt.gz; Human ENSEMBL annotations file.
http://genome.cse.ucsc.edu; UCSC genome browser.
http://www.ensembl.org; ENSEMBL genome browser.
http://www1.imim.es/genepredictions/; GENEID and SGP2 full data predictions.
http://www1.imim.es/software/sgp2/; SGP2 home page.
http://www1.imim.es/datasets/sgp2002/; SGP2 training data sets page.

Received November 4, 2002; accepted in revised form November 15, 2002.

4.4 Accuracy of gene prediction methods

In collaboration with the genome sequencing centers, the Vertebrate Genome Annotation database (VEGA, `http://vega.sanger.ac.uk/`) attempted to present a consistent high-quality curation of vertebrate genomic sequences. Finished genomic sequences are analyzed on a clone by clone basis using a combination of similarity searches against DNA and protein databases as well as a series of *ab initio* gene predictions. The data gathered in these steps is then used to manually annotate each gene structure. The annotation is based on supporting evidence only. However, VEGA uses `genscan` and `fgenesh` in the annotation pipeline, and may be biased toward these programs. Currently, there are eight manually annotated human chromosomes.

We decided to evaluate the accuracy of a number of *ab initio* and comparative gene finders in chromosome 22 again. This time using the curated annotations from VEGA as reference. The results that we obtained, shown in Table 4.1, are consistent with the ones presented in the Parra et al. (2003). Sensitivity and specificity were a little bit higher because extra genes were added to the initial chromosome 22 annotation set used in the previous paper.

As we have already seen, accuracy suffers substantially when moving from single gene sequences to whole chromosome sequences. For instance, `genscan` *CC* drops from 0.91 in the evaluation by Rogic et al. (2001) (shown in Figure 1.3) to 0.65 for chromosome 22.

Ab initio programs have a similar accuracy being `geneid` more specific and `genscan` more sensitive. Both programs tend to predict short wrong genes that do not overlap with any real gene corresponding nearly half of the predictions to wrong genes.

Comparative approaches show an important improvement in comparison with the *Ab initio* methods, specially in specificity. *Ab initio* specificity at nucleotide level is on average 0.56, while comparative methods reach a specificity of 0.70. Sensitivity at exon and gene level increase almost 10% on average. `sgp2` performs better at nucleotide level while `twinscan` at exon level. In spite of the general improvement of comparative gene prediction, still 30% of predicted genes, on average, do not overlap any real annotation.

	Base level			Exon level					Gene level				
	Sn	Sp	CC	Sn	Sp	SnSp	ME	WE	Sn	Sp	SnSp	MG	WG
ab initio													
`genscan`	0.88	0.49	0.65	0.72	0.40	0.56	0.12	0.51	0.07	0.04	0.05	0.12	0.53
`geneid`	0.82	0.63	0.71	0.70	0.54	0.62	0.17	0.36	0.13	0.08	0.10	0.13	0.46
comparative													
`twinscan`	0.83	0.69	0.74	0.77	0.63	0.70	0.11	0.26	0.16	0.14	0.15	0.17	0.24
`sgp2`	0.85	0.70	0.76	0.74	0.60	0.67	0.10	0.31	0.15	0.09	0.12	0.13	0.33
sequence similarity based													
`fgenesh`	0.94	0.68	0.79	0.90	0.67	0.78	0.05	0.29	0.58	0.35	0.47	0.09	0.42
ENSEMBL	0.82	0.82	0.81	0.73	0.76	0.75	0.19	0.15	0.43	0.34	0.38	0.09	0.22

Table 4.1: Accuracy of different gene finding tools on the human chromosome 22 using as reference the VEGA annotations.

Even more sophisticated annotation pipelines, such as ENSEMBL (based on `genewise`) or `fgenesh`, which use known cDNAs, are far from producing perfect predictions, with *CC*s around 0.80 and prediction at gene level around 0.40. These numbers strongly suggest that current mammalian gene counts are still of a highly hypothetical nature.

4.5 sgp2 distribution and web server

The `sgp2` distribution contains a set of independent programs written in C and PERL. `sgp2` is distributed through the GNU General Public License (GNU GPL, `http://www.gnu.org/`). That means that the code is freely available to any user. GNU GPL gives the legal permission to copy, distribute and/or modify the software.

We have also developed a web server (see Figure 4.4) to allow the usage through the Internet. The available web server is optimized for human-mouse comparative prediction, but our plans include the extension to other species in the future.

`sgp2` was extensively used in the annotation of the mouse and the human genome. In the *Annexed paper* section is attached the paper of the Mouse Genome Consortium in which we have participate obtaining and processing the set of `sgp2` predictions.

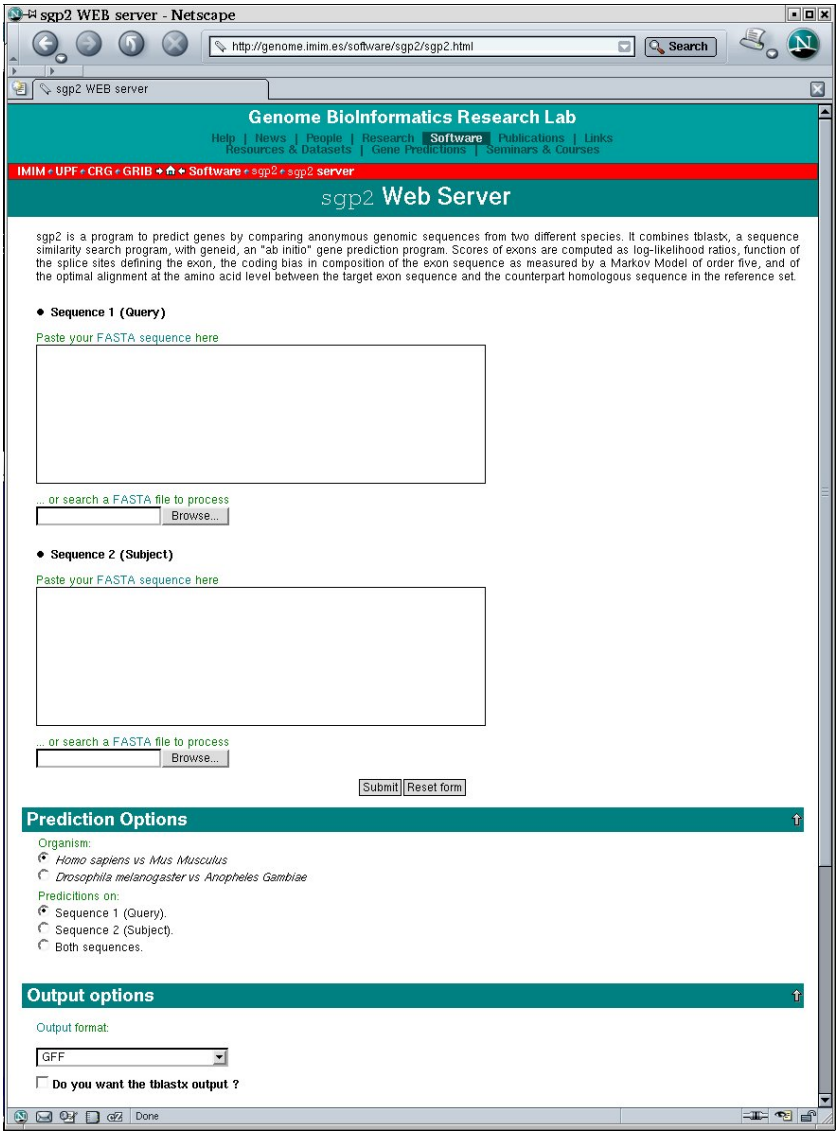


Figure 4.4: Form of the `sgp2` web interface server accesible through `http://genome.imim.es/software/sgp2/sgp2.html`.

Toward the completion of the mammalian catalog of genes

The completion of the mouse genome allowed for the first time a comparative based annotation of two mammalian genomes and several comparative methods were applied to improve gene predictions. However, current predictions are not reliable enough. This section describes a protocol that reduces the false positive rate of predictions by exploiting the exonic conservation between human and mouse homologous genes. Using this protocol, a set of human-mouse predicted genes was generated and partially validated by experimental approaches.

5.1 Expanding Human and Mouse standard annotation pipelines

The challenge of predicting coding genes in genome sequences has been broadly discussed in previous sections. Despite the improvements achieved using comparative approaches, gene prediction methods still tend to predict many false positives. Therefore, for the initial annotation of the mouse genome, the Mouse Genome Sequencing Consortium relied mainly on the ENSEMBL gene build pipeline. The ENSEMBL automatic annotation pipeline basically relies on known proteins and mRNA sequences (as explained in section 1.3). However, ENSEMBL can not predict genes for which there is no preexisting evidence of transcription or similarity to a closely related protein. This limitation could lead to a bias against predicting genes that have a restricted expression pattern.

The aim of the following work was to survey how many coding genes have been missed by the conservative ENSEMBL gene building pipeline. The goal, was to generate a set of novel predicted genes that was not included in the ENSEMBL annotation. To achieve a high standard of accurate predictions, a filtering protocol based on the exonic structure conservation of orthologous human-mouse genes was developed. Samples of gene predictions from each step were validated experimentally to assess the efficiency of the protocol.

`sgp2` and `twinscan` were used to generate the initial set of gene predictions. Both programs are based on comparative gene prediction approaches, and both clearly out-

perform *ab initio* gene prediction programs (as showed in Table 4.1, Korf et al. (2001) and Parra et al. (2003)).

The protocol, performed independently within each predictor, consisted of:

- the prediction of human and mouse genes using both comparative gene predictors: `sgp2` and `twinscan`.
- the identification of homologous human-mouse prediction pairs based on protein similarity using `blastp`.
- the filtering of predictions that overlap ENSEMBL predictions.
- the generation of a set of pairs of homologous predictions, with conserved exonic structure.

After the different enriched sets of gene predictions were generated, experimental validation of samples of each group was obtained by RT-PCR and direct sequencing.

Our contribution to this project was the generation of the `sgp2` predictions in the human and in the mouse genome and the development of the filtering process that leads to the different subsets of gene predictions. We also contributed to the development of `exstral`, the program needed for the superimposition of the exon-intron boundaries over a protein alignment.

5.2 Obtaining sgp2 predictions

In this analysis, predictions have been obtained on the mouse genome (MGSCv3 assembly) using comparative information from the human genome (NCBI Build 28) and viceversa. To obtain the similarity regions, mouse chromosomal sequences were split into 100 kb fragments to build the `blast` database. The masked human chromosomes were also split in 100 kb fragments which were compared with the mouse database using `tblastx` with parameters to speed up the search (as explained in Parra et al. (2003)). Although these parameters increased the speed of the comparison, the whole computation took one week of CPU time using 100 Alpha processors.

The 7,194,658 HSPs resulting from the comparison of the human and the mouse genomes were processed in order to find the maximum scoring projections (MSPs). The MSPs correspond to the complete non overlapping sections of HSPs with the highest score. This process is realized taking into account in which of the six coding frames the HSPs have been found (as far as the alignments provided by `tblastx` are obtained at protein level). After the projection, the number of HSPs was reduced to 2,169,704 MSPs for human and 2,145,493 for mouse.

`sgp2` has essentially no limits to the length of the input query sequence, and deals well with chromosome sequences. Therefore, predictions were computed from the entire chromosome sequences (no fragmentation was needed). The predictions were obtained from the unmasked sequences of the human and the mouse genome. The computation took one day in a MOSIX cluster containing four PCs (PentiumIII Dual 500 Mhz processors).

sgp2 predictions were obtained in a mode in which the complete mRNA sequences obtained from the Reference Sequence database (REFSEQ, http://www.ncbi.nlm.nih.gov/RefSeq/), mapped by the UCSC browser team, were provided to sgp2 as external gene evidence. The coordinates of the corresponding coding fraction of the mRNAs were provided to sgp2 in GFF format. Thus, predictions were built on top of this experimentally verified set of mRNAs. Obviously, these genes were correctly predicted, but in addition, the incorporation of these genes as external information, induced sgp2 to refine the structure of nearby genes, reducing the number of joined and split genes. This is a serious problem of most gene finding programs, because of the poor conservation of the signals defining the beginning and the end of genes. Using this external information, sgp2 only predicted genes in the regions between known mRNAs.

The accuracy of the predictions using external evidence was evaluated in the human chromosome 22. The evaluation of the accuracy using VEGA annotations as reference is shown in Table 5.1. The REFSEQ mRNA set (510 transcripts corresponding to 380 genes) can be considered as a subset of the VEGA annotation set (containing 493 genes). The relatively low specificity at gene level obtained by the REFSEQ set, 0.76, can be explained by the fact that only one of the overlapping transcripts per gene is taken into account for the evaluation. Therefore, in some cases, the selected isoform obtained from the VEGA annotations does not correspond to the same isoform of the REFSEQ set. The sensitivity of sgp2 at nucleotide level using the REFSEQ set increased only from 0.84 to 0.94. However, the sensitivity at gene level increased from 0.15 to 0.68. The number of genes predicted by sgp2 without using the mRNAs set in human chromosome 22 was 711 and using the set of 380 non-overlapping mRNAs was 727.

	Base level			Exon level					Gene level		
	Sn	Sp	CC	Sn	Sp	Sn+Sp	ME	WE	Sn	Sp	SnSp
REFSEQ mRNA	0.86	0.96	0.91	0.88	0.95	0.91	0.10	0.02	0.66	0.76	0.71
sgp2	0.85	0.70	0.76	0.74	0.60	0.67	0.10	0.31	0.15	0.09	0.12
sgp2 mRNA	0.94	0.74	0.84	0.92	0.70	0.81	0.04	0.20	0.68	0.38	0.53

Table 5.1: Accuracy of sgp2 on human chromosome 22 using REFSEQ mRNAs as external evidence and the VEGA annotations as reference. The first sgp2 row contains the accuracy of the standard sgp2. In the sgp2 mRNA the results of introducing the mRNAs of the REFSEQ database.

Figure 5.1 shows the schema of the sgp2 prediction protocol. sgp2 predicted 45,104 genes (including the 14,729 human mRNAs obtained from REFSEQ database) in the human genome and 47,055 genes (including the 8,405 mouse mRNAs from REFSEQ database) in the mouse genome.

twinscan predictions were provided directly by Michael Brent's laboratory at Washington University (St. Louis) and they were incorporated in the next steps of the protocol.

5.3 Obtaining the homologous pairs of predictions

The enrichment procedure was applied separately to twinscan and sgp2 predictions. The protein sequences predicted by each program in the mouse genome were compared

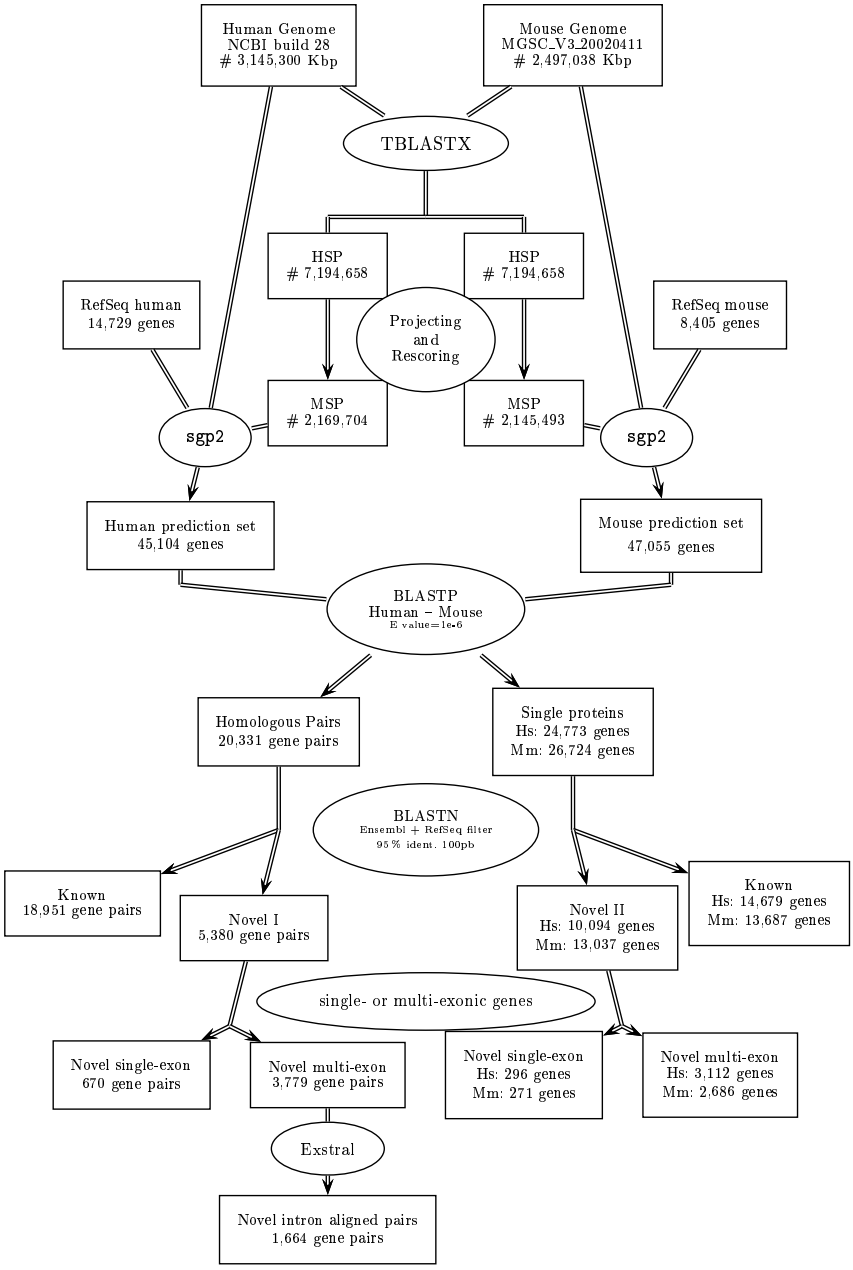


Figure 5.1: Schema of the protocol to obtain human-mouse `sgp2` prediction and filtering process

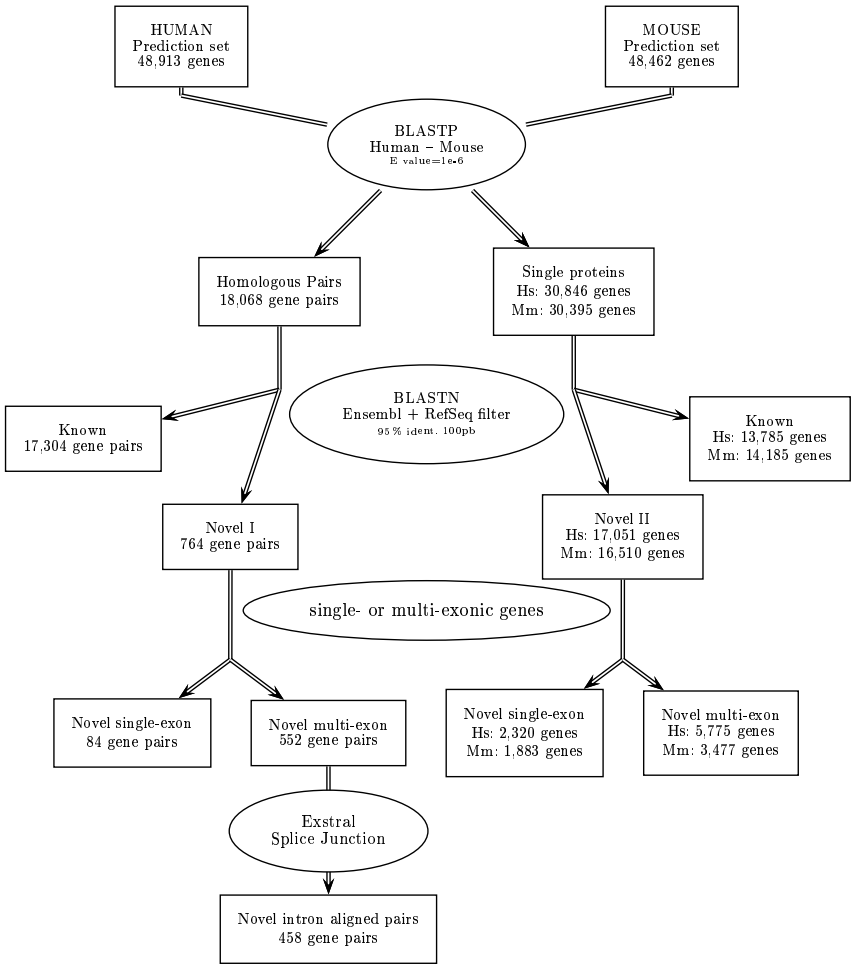


Figure 5.2: Schema of the filtering process for `twinscan` comparative human-mouse predictions

with the human predictions using `blastp`. For each predicted mouse protein, human predictions with `blast` expected values (e-value) lower than $1x10^{-6}$ were considered homologs. 18,068 and 20,331 homologous human-mouse pairs of predictions were obtained from `twinscan` and `sgp2` respectively.

As far as the purpose of the experiment was to find novel genes, the predictions corresponding to known genes were discarded. The preliminary ENSEMBL annotation generated with the RIKEN cDNA database was used as the standard of known genes. Any prediction that overlapped with the mapped ENSEMBL genes was rejected (considered non novel). Moreover, to assure the novelty, predictions were compared with the ENSEMBL predicted transcripts and the complete REFSEQ database using `blastn`. Predictions with more than 95% nucleotide identity over at least 100 bp were also rejected. The novelty protocol was applied to the mouse and human gene predictions. In the case of the homologous predictions, if only one of the pair of proteins was found in the set of known proteins both were discarded. About 95% of the homologous set of predicted proteins were considered to be previously known (18,951 of 20,331 `sgp2` and 17,304 of 18,068 `twinscan` predictions). However, only about 50% of the singleton proteins correspond to the set of known proteins (13,687 of 26,724 `sgp2` and 14,185 of 30,395 predictions).

This low percentage of known proteins in predictions not having a homologous counterpart suggested that this set could not be very reliable. We analyzed some of these cases and we found that many genes were shorter than expected and contained highly repetitive stretches of DNA. These low complexity regions can mislead measurements of coding statistics due to their repetitive composition, and may be scored as coding regions. A quality filter was applied to this set in order to obtain a better set of predictions. Genes shorter than 100 bp or those of which more than 75% of the prediction corresponded to low complexity regions were discarded. The `dust` program (included in the `WU-blast` support programs, `http://blast.wustl.edu/`) was used to determine the percentage of low complexity regions in the predicted transcripts. `dust` detects highly repetitive regions, variable number of tandem repeats and short tandem repeats. Of the set of 13,037 `sgp2` and 16,510 `twinscan` unpaired proteins 77,31% and 78,59% were respectively discarded after the quality control filtering.

5.4 Conserved exonic structure

Mutations that disrupt splicing can cause catastrophic reading frame shifts, therefore intron junctions and exonic structure are expected to be conserved features. A complete analysis of conserved exonic structure is described by the Mouse Genome Sequencing Consortium (2002) from a set of 1,506 pairs of human-mouse REFSEQ genes confidently assigned to be orthologous. The Mouse Genome Sequencing Consortium (2002) showed that gene structures are very conserved between orthologous pairs: 86% of the cases have the identical number of coding exons and 46% have identical coding sequence length. When all exons, rather than just coding exons, are taken into account, 62% have the same number of exons. Based on this data we have developed a method to check if pairs of genes from the homologous set of predictions have a conserved exonic structure.

Every homologous pair of predictions was first aligned using `t-coffee` (Notredame et al., 2000). `t-coffee`, a global sequence alignment program, was run with default pa-

rameters on the amino acid sequences. Exonic structure was added to the global pairwise alignments using `exstral` (Exon Structural Alignment).

`exstral` is a program that takes a global alignment of two proteins and the genomic coordinates of the exonic structure of both genes as input, and outputs the exon and intron junctions superimposed on the protein alignment. This program computes the relative position of the intron boundaries in aligned pairs of sequences. The exon-intron junctions are superimposed on the alignment taking into account the corresponding position of the amino acid as well as the codon position where the exon-intron junction occurs. In addition of the alignment with the exonic structure,`exstral` also provides information of each confirmed compatible exonic junctions. Figure 5.3 shows the output of `exstral`.

When both members of an aligned gene pair contained an intron at the same coordinate with at least 50% identity over 15 amino acids at both sides of the alignment, it was assigned to the "enriched" pool. Predictions with homologous proteins but no aligned introns were assigned to the "similar" pool.

5.5 RT-PCR validation experiments

A subset of random predictions were extracted from each set (for `sgp2` and `twinscan`), and two adjacent exons across an intron were chosen from the selected predictions for the RT-PCR test. The experimental test required that the exons were at least 30 bp long, and the introns were at least 1000 bp long. Pairs of exons verifying these requirements are sorted by the sum of the scores given by each prediction program, and the top scoring pair was selected for the RT-PCR test.

5.6 Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes.

The following article is mainly based on the filtering method and the different subsets of predictions we have already described. The final results of the RT-PCR experiments showed that the comparative enrichment selection correlates with the ratio of amplification. It also contains a functional analysis of some of the predicted proteins

```
chr12_328       MSVTGFTITDEKVHLYHSIEKEKTVRHIGDLCSSHSVKKIQVGICLLLVELCERFTFFEV

chr6_2206       ------------------------------------------------------------


chr12_328       VCNMIPFCTIKLGYHNCQAAILNLCFIGTSILTPVFVRWLTDVYLGRNKLVYICLFLHFL
                   ****** :** .* ***:*** *****:*****: **:* *:*****:**.* ****
chr6_2206       ---MIPFCTGRLGSYNHQAAMLNLGFIGTSVLTPVFMGWLADEYFGRNKLMYIALSLHFL


                |1a
chr12_328       GTALLSVVAFPLEDFYLGTYHAVNNIPKTEQHRLFYVALLTICLGIGGVRAIVCPLGAFG
                ******:.*** *:** *:* ..** .  **  **:*****:*** **:**:*** .  *
chr6_2206       GTALLSMLAFPAENFYRGAYPVFNNTSVEEQAGLFHVALLTLCLGTGGIRAVVCPPDMCG
                |1a

                          |2b
chr12_328       LQEYGSQKTMSFFNWFYWLMNLNATIVFLGISYIQHSQAWALVLLIPFMSMLMAVITLHM
                 **  *:*.*.* **  *  ****:*****  **   : ** :*:* :*:: *::**::
chr6_2206       SQERESKKPMPFCNWASWSANLNAAVVFLGISSIQPLGSGALGILLPSLSVFTALVTLYL
                          |2b

                       |3a
chr12_328       IYYNLIYQSEKRVGVL------VSALKTCHPQYCHLGRDVTSQLDHAKEKNGGCYSELHV
                 : :***:.*:* .:*      * ****    ***:*** .* **** **:** :***:
chr6_2206       KHCDLIYRPENRCSLLTIARAFVRALKTRCLPYCHFGRDGSSWLDHAMEKQGGHHSELQE
                       |3a

                                  |4c
chr12_328       EDTTFFLTLLPLFIFQLLYRMCIMQIPSGYYLQTMNSNLNLDGFLLPIAVMNAISSLPLL
                ***  : :***** **:*** *::************** * .** ****:***** ****
chr6_2206       EDTRNISALLPLFSFQILYRTCLLQIPSGYYLQTMNSNRNWGGFSLPIALMNAISLLPLL
                                  |4c

                                     |5a
chr12_328       ILAPFLEYFSTCLFPSKRVGSFLSTCIIAGNLFAALSVMIAGFFEIHRKHFPAVEQPLSG
                **.**::***.**:**** *.***:*:****: ** ** :***:**:**   * **. **
chr6_2206       ILPPFMDYFSNCLLPSKRDGPFLSACMIAGNICAASSVAMAGFLEIYRK--LAREQSPSG
                                     |5a

                                        |6b
chr12_328       KVLTVSSMPCFYLILQYVLLGVAETLVNPAREYCNLN----------------------
                *:::****.*. *: *******:*.*****   *. .
chr6_2206       KLFSVSSMACVCLVPQYVLLGVSEVLVNPAGAQCSKTCGIIKYSQKLATLQLSDNRTFAP
                                        |6a


chr12_328       -------HFNA--------------------------QNIRGSNLEETLLLHEKSLKFYG
                       *:.                        :.***.. ****** ********
chr6_2206       NRQLVSKHIRQRQSRELLLRAGILDAAQHPGTLEGFMERIRGNRCEETLLLTEKSLKFYG
                                                 |7c


chr12_328       SIQEFSSSIDLWETAL
                * *  ***********
chr6_2206       STQGASSSIDLWETAL


chr12_328 chr6_2206 1 1 121a 361 223 172 223   0.94 0.93   0.81 0.73   0.00 0.00

chr12_328 chr6_2206 2 2 195b 223 170 223 170   0.75 0.73   0.56 0.60   0.00 0.00

chr12_328 chr6_2206 3 3 252a 170 203 170 221   0.88 0.47   0.44 0.33   0.00 0.40

chr12_328 chr6_2206 4 4 325c 203 187 221 187   0.88 0.93   0.69 0.93   0.00 0.00

chr12_328 chr6_2206 5 5 388a 187 193 187 183   0.94 0.80   0.69 0.67   0.00 0.00

# chr12_328 chr6_2206       6       7       5       0.48    0.63    0.83
```

Figure 5.3: `exstral` alignment output. On top, the alignment of the two proteins with the superimposition of the exonic boundaries. Each exon boundary is marked with a vertical line, the number of intron, and the codon position (a, b or c). The bottom part of the `exstral` output shows the information associated with the aligned boundaries. For each aligned exon-exon boundary the names of the proteins, the number of the aligned exons, the length of the upstream and downstream exons, and the percentage of identities, similarities and gaps of the the 15 downstream and upstream amino acid are shown.

Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes

Roderic Guigó*†, Emmanouil T. Dermitzakis†‡, Pankaj Agarwal§, Chris P. Ponting¶, Genís Parra*, Alexandre Reymond‡, Josep F. Abril*, Evan Keibler‖, Robert Lyle‡, Catherine Ucla‡, Stylianos E. Antonarakis‡, and Michael R. Brent‖**

*Research Group in Biomedical Informatics, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra/Centre de Regulació Genòmica, E08003 Barcelona, Catalonia, Spain; ‡Division of Medical Genetics, University of Geneva Medical School and University Hospitals, 1211 Geneva, Switzerland; §GlaxoSmithKline, UW2230, 709 Swedeland Road, King of Prussia, PA 19406; ¶Medical Research Council Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, United Kingdom; and ‖Department of Computer Science, Washington University, One Brookings Drive, St. Louis, MO 63130

Communicated by Robert H. Waterston, Washington University School of Medicine, St. Louis, MO, December 11, 2002 (received for review October 21, 2002)

A primary motivation for sequencing the mouse genome was to accelerate the discovery of mammalian genes by using sequence conservation between mouse and human to identify coding exons. Achieving this goal proved challenging because of the large proportion of the mouse and human genomes that is apparently conserved but apparently does not code for protein. We developed a two-stage procedure that exploits the mouse and human genome sequences to produce a set of genes with a much higher rate of experimental verification than previously reported prediction methods. RT-PCR amplification and direct sequencing applied to an initial sample of mouse predictions that do not overlap previously known genes verified the regions flanking one intron in 139 predictions, with verification rates reaching 76%. On average, the confirmed predictions show more restricted expression patterns than the mouse orthologs of known human genes, and two-thirds lack homologs in fish genomes, demonstrating the sensitivity of this dual-genome approach to hard-to-find genes. We verified 112 previously unknown homologs of known proteins, including two homeobox proteins relevant to developmental biology, an aquaporin, and a homolog of dystrophin. We estimate that transcription and splicing can be verified for >1,000 gene predictions identified by this method that do not overlap known genes. This is likely to constitute a significant fraction of the previously unknown, multiexon mammalian genes.

Complete and precise delineation of protein coding genes in mammalian genomes remains a challenging task. To produce a preliminary gene catalog for the draft sequence of the mouse (1), the Mouse Genome Sequencing Consortium relied primarily on the ENSEMBL gene build pipeline (2). ENSEMBL works by (*i*) aligning known mouse cDNAs from REFSEQ (3), RIKEN (4, 5), and SWISSPROT (6, 7) to the genome, (*ii*) aligning known proteins from related mammalian genes to the genome, and (*iii*) using portions of GENSCAN (8) predictions that are supported by experimental evidence (such as ESTs). This conservative approach yielded ≈23,600 genes. However, ENSEMBL cannot predict genes for which there is no preexisting evidence of transcription (1). Furthermore, reliance on known transcripts may lead to a bias against predicting genes that are expressed in a restricted manner or at very low levels.

Before the production of a draft genome sequence for a second mammal, the best available methods for predicting novel mammalian genes were single-genome *de novo* gene-prediction programs, of which GENSCAN (8) is one of the most accurate and most widely used. These programs work by recognizing statistical patterns characteristic of coding sequences, splice signals, and other features in the genome to be annotated. However, they tend to predict many apparently false exons caused by the occurrence of such patterns by chance. With the availability of draft sequences for both the mouse and human genomes, it is now possible to incorporate genomic sequence conservation into *de novo* gene prediction algorithms. However, DNA alignment programs alone are not an effective means of gene prediction because a large fraction of the mouse and human genomes is conserved but does not code for protein.

We developed a procedure that greatly reduces the false-positive rate of *de novo* mammalian gene prediction by exploiting mouse–human conservation in both an initial gene-prediction stage and an enrichment stage. The first stage is to run gene-prediction programs that use genome alignment in combination with statistical patterns in the DNA sequence itself. A number of such programs have been described (9–12). For these experiments, we used SGP2 (13) and TWINSCAN (refs. 14 and 15 and http://genes.cs.wustl.edu), two such programs that we designed for efficient analysis of whole mammalian genomes. TWINSCAN is an independently developed extension of the GENSCAN probability model, whereas SGP2 is an extension of GENEID (16, 17). The probability scores these programs assign to each potential exon are modified by the presence and quality of genome alignments. TWINSCAN uses nucleotide alignment [BLASTN (18), blast.wustl.edu] and has specific models for how alignments modify the scores of coding regions, UTRs, splice sites, and translation initiation and termination signals. SGP2, in contrast, uses translated alignments [TBLASTX (18), blast.wustl.edu] to modify the scores of potential coding regions only. These programs predict many fewer exons than GENSCAN with no reduction in sensitivity to the exons of known genes (13, 14).

The second stage of our procedure is based on the observation that almost all mouse genes have a human counterpart with highly conserved exonic structure (1). We therefore compare all multiexon genes predicted in mouse in the first stage to those predicted in human. Predictions are retained only if the protein predicted in mouse aligns to a human protein predicted by the same program, with at least one predicted intron at the same location (aligned intron, Fig. 1). Predicted single-exon genes are always discarded by this procedure. Although there are many real single-exon genes, it is not currently possible to predict them reliably nor to verify them reliably in a cost-effective, high-throughput procedure.

In this article, we show that our two-stage process yields >1,400 predictions outside the standard annotation of the mouse genome. RT-PCR and direct sequencing of a single exon pair in a sample of these predictions indicates that the majority correspond to real spliced transcripts. Our results also show that this procedure is sensitive to genes that are hard to find by other methods. The combination of these computational and experimental techniques forms a powerful, cost-effective system for expanding experimentally supported genome annotation. This approach is therefore expected to bring the annotation of the mouse and human genomes nearer to closure.

Experimental Procedures

Genome Sequences. The MGSCv3 assembly of the mouse genome described in ref. 1 and the December, 2001 Golden Path assembly

†R.G. and E.T.D. contributed equally to this work.

**To whom correspondence should be addressed. E-mail: brent@cse.wustl.edu.

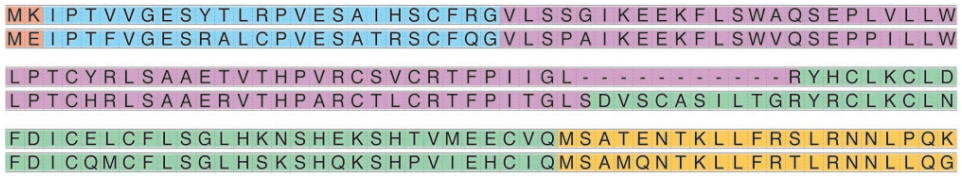


Fig. 1. An example of predictions with aligned introns. RT-PCR positive predicted protein 3B1 (a novel homolog of *Dystrophin*) is aligned with its predicted human ortholog (N-terminal regions shown; *Upper* of each row: mouse, *Lower* of each row: human). Each color indicates one coding exon. Three of four predicted splice boundaries (color boundaries) align perfectly. Any one of these three is sufficient for surviving the enrichment step. Gaps in the alignment (shown as dashes) may indicate mispredicted regions.

of the human genome (National Center for Biotechnology Information Build 28) were downloaded from the University of California (Santa Cruz) genome browser (http://genome.ucsc.edu).

Genome Alignments. TWINSCAN was run on the mouse genome by using BLASTN alignments to the human genome (WU-BLAST, http://blast.wustl.edu). Lowercase masking in the human sequence was first converted to N masking. The result was further masked with NSEG by using default parameters, all Ns were removed, and the sequence was cut into 150-kb database segments. The mouse genome sequence was divided into 1-mb query segments. BLASTN parameters were: M=1 N=−1Q=5 R=1 Z=3000000000 Y=3000000000 B=10000 V=100 W=8 X=20 S=15 S2=15 gapS2=30 lcmask wordmask=seg wordmask=dust topcomboN=3. TWINSCAN was run on the human genome by using separate BLASTN alignments to the mouse genome, which was prepared in the same way except that Ns were not removed before creating the BLAST database.

SGP2 was run on the mouse and human genomes by using a single set of alignments. The masked human genome was cut into 100-kb query segments that were compared with a database of all 100-kb segments of the mouse genome with TBLASTX (WU-BLAST, parameters: B=9000 V=9000 hspmax=500 topcomboN=100 W=5 E=0.01 E2=0.01 Z=3000000000 nogap filter=xnu+seg S2=80). The substitution matrix was BLOSUM62 modified to penalize alignments with stop codons heavily (−500).

Initial Gene Predictions. TWINSCAN was run on 1-mb segments of the mouse and human genomes with target genome parameters identical to the GENSCAN parameters and the 68-set-ortholog conservation parameters (available on request). Note that the TWINSCAN results described in ref. 14 are based on a subsequently developed set of target genome parameters that yields better results than those described here. SGP2 was run on unsegmented mouse and human chromosomes. The REFSEQ genes (which were not tested in the experiments reported here) were incorporated directly into the SGP2 predictions, which improved the predictions outside the REFSEQS slightly by preventing some gene fusion errors. Note that the REFSEQS were not used in generating the SGP2 results described in ref. 13.

Novelty Criteria. Mouse predictions were considered known if they overlapped ENSEMBL predictions or had 95% nucleotide identity to a REFSEQ mRNA or an ENSEMBL-predicted mRNA over at least 100 bp. We used the most inclusive set of ENSEMBL predictions available, based on the complete RIKEN cDNA set without further filtering (1).

Enrichment Procedure. The enrichment procedure was applied separately to predictions of TWINSCAN and SGP2. The protein sequences predicted by each program in human and mouse were compared by using BLASTP (19). For each predicted mouse protein, all predicted human proteins with expect values $<1 \times 10^{-6}$ were called homologs. A global protein alignment was produced for the best scoring homologs (up to five) by using T-COFFEE (ref. 39; http://igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html) with default parameters. Exonic structure was added to the alignments by using EXSTRAL.PL (www1.imim.es/~rcastelo/exstral.html). When both members of an aligned pair contained an intron at the same coordinate with at least 50% identity over 15 aa on both sides the corresponding mouse prediction was assigned to the "enriched" pool. Predictions with homologs but no aligned intron were assigned to the "similar" pool.

RT-PCR. To test predictions, primers were designed in adjacent exons as described in *Results* and used in RT-PCR of total RNA from 12 normal mouse adult tissues. All procedures were as described (20), except that JumpStart REDTaq ReadyMix (Sigma) and primers from Sigma-Genosys were used.

Additional Details. See supplementary information at www1.imim.es/datasets/mouse2002 for additional details of these procedures.

Results

We applied the two-stage procedure described above to the entire draft mouse and human genome sequences (see *Experimental Procedures*). TWINSCAN predicted 17,271 genes with at least one aligned intron, whereas SGP2 predicted a largely overlapping set of 18,056 genes with at least one aligned intron. These predicted gene sets contain 145,734 exons and 168,492 exons, respectively. Together the two sets overlapped 90% of multiexon ENSEMBL gene predictions.

To estimate a lower bound on the proportion of novel predictions that are transcribed and spliced, we performed a series of RT-PCR amplifications from 12 adult mouse tissues (20). We did not test genes that overlap ENSEMBL predictions nor those that are 95% identical to ENSEMBL predictions or REFSEQ mRNAs over >100 bp or more. Because ENSEMBL was the standard for annotation of the draft mouse genome, we refer to the non-ENSEMBL genes as "novel." A random sample of novel genes predicted by each program and containing at least one aligned intron was tested. Primer pairs were designed in adjacent exons separated by an aligned intron of at least 1,000 bp (Fig. 2). The exon pair to be tested was chosen on the basis of intron length (minimum 1,000 bp), primer design requirements, and *de novo* gene prediction score, with no reference to protein, EST, or cDNA databases. Amplification followed by direct sequencing of the PCR product (Fig. 3) verified the exon pair in 133 unique predicted genes of 214 tested (62%, enriched pool, see Table 1 and www1.imim.es/datasets/mouse2002). Mouse genes predicted by both programs were verified at a much higher rate than those predicted by just one program (76% vs. 27%). Extrapolating from the success rates in Table 1, testing the entire pool of 1,428 enriched predictions in this way is

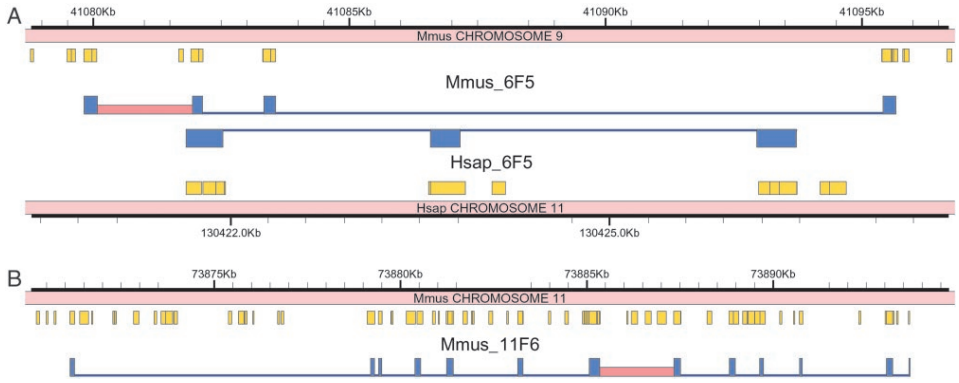


Fig. 2. Two examples of predicted gene structures (blue) with introns verified by RT-PCR from primers located in exons flanking the introns indicated in red. Mouse–human genomic alignments (orange) correlate with predicted exons but do not match them exactly. (*A*) Verified mouse prediction 6F5, a novel homolog of *Drosophila* brain-specific homeobox protein (bsh), with matching human prediction. (*B*) Verified mouse prediction 11F6, a homolog of rat vanilloid receptor type 1-like protein 1. No matching human gene was predicted. A cDNA (GenBank accession no. AF510316) that matches the predicted protein over four protein-coding exons was deposited in GenBank subsequent to our analysis.

expected to yield a total of 788 (±48) predictions with confirmed splices, none of which overlap ENSEMBL predictions.

Considered in isolation, genes predicted by TWINSCAN had a higher verification rate than those predicted by SGP2 (83% vs. 44%), but that difference is skewed by the fact that TWINSCAN predicted fewer exons per gene, and hence its predictions were less likely to overlap ENSEMBL predictions. We corrected for this by clustering overlapping TWINSCAN and SGP2 predictions to ensure that both were counted as positive if either was verified experimentally. For each program, the predictions belonging to a given cluster were counted only once, even if more than one was RT-PCR positive. After this correction, the confirmation rates were much closer (76% for TWINSCAN vs. 62% for SGP2). The results shown in Table 1 include the correction. The TWINSCAN verification rate is similar to the verification rate for genes predicted by both programs because the exons predicted by TWINSCAN are largely a subset of those predicted by SGP2.

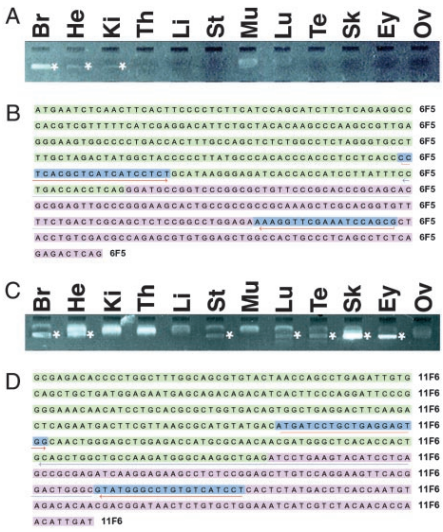


Fig. 3. Verification of gene predictions by RT-PCR analysis. (*A* and *B*) Test of prediction 6F5, a homolog of *Drosophila* brain-specific homeobox protein (bsh). (*C* and *D*) Test of prediction 11F6, a homolog of rat vanilloid receptor type 1-like protein. Gel analysis of amplimers (*) with the source of the cDNA pool indicated above is shown in *A* and *C*. Primers (blue) and the region to which the amplimer sequence aligned (underlining) are shown in *B* and *D*. The indicated forward primers were used to generate the amplimer sequences (brain amplimer, *B*; skin amplimer, *D*). Br, brain; Ey, eye; He, heart; Ki, kidney; Li, liver; Lu, lung; Mu, muscle; Ov, ovary; Sk, skin; St, stomach; Te, testis; Th, thymus.

Before the enrichment procedure, the combined predictions of SGP2 and TWINSCAN overlap 98% of multiexon ENSEMBL genes, as compared with 90% for the enriched pool. This finding suggests that the enrichment procedure reduces sensitivity by a small but noticeable degree. To investigate the potential loss of sensitivity further, we applied the same RT-PCR procedure to two samples of gene predictions that were excluded by the enrichment criterion and did not overlap ENSEMBL predictions. One sample had one or more regions of strong similarity to a predicted human gene but did not satisfy the aligned intron criterion (similar pool) whereas the other lacked any strong similarity to a human prediction by the same program (other pool). The verification rates for the similar and other pools were 25% and 20%, respectively, for genes predicted by both programs, and 0% and 2%, respectively, for genes predicted by only one program (Table 1 and www1.imim.es/datasets/mouse2002). This finding shows that the enrichment procedure increases specificity greatly and, consistent with the ENSEMBL overlap analysis, reduces sensitivity only slightly. If all predictions in the similar and other pools were tested the expected numbers of successes are 126 (±105) and 105 (±83), respectively, with the large standard errors resulting from the small number of successful amplifications in these pools.

As a control, we also tested 113 predictions from the enriched pool that did overlap ENSEMBL predictions. In 66 of the predictions the splice boundary we tested was predicted identically in ENSEMBL, and 64 of these tests (97%) were positive. In 47 of the predictions the splice boundary we tested was not predicted identically in ENSEMBL, and 21 of these tests (45%) were positive,

Table 1. Predicted novel gene sets and RT-PCR verification rates

Pool	Programs*	No. of predictions	No. tested	No. positive	Success rate, %	Expected successes	Standard error
Enriched†	Both	827	154	117	75.97	628	
	One	601	60	16	26.67	160	
	Total	1,428	214	133	62.15	788	48
Similar‡	Both	505	16	4	25.00	126	
	One	1,620	22	0	0.00	0	
	Total	2,125	38	4	10.53	126	105
Other§	Both	234	5	1	20.00	46	
	One	3,425	58	1	1.72	59	
	Total	3,659	63	2	3.17	105	83
All	Total	7,212	315	139	N/A	1,019	

N/A, not applicable.

*Both, Genes predicted at least partially by both TWINSCAN and SGP2 programs. One, Genes predicted by one program that are not overlapped by predictions of the other program. N/A, not applicable.

†Mouse gene predictions containing an intron whose flanking exonic regions align with flanking exonic regions predicted by the same program in human.

‡Mouse gene predictions that fail the enrichment step but show regions of strong similarity to a gene predicted by the same program in human.

§Mouse gene predictions without regions of strong similarity to any gene predicted by the same program in human.

despite the fact that ENSEMBL predictions are based on transcript evidence. This verification rate may reflect alternative splices identified by our method but not by ENSEMBL.

To determine whether tissue-restricted expression could explain the absence of the predictions we verified from the transcript-based annotation, we compared the expression patterns of our RT-PCR positive predictions to those of the complete set of mouse orthologs of genes mapping to human chromosome 21 (Hsa21). These genes were chosen for comparison because they had been previously subjected to the same protocol with the same cDNA pools in the same laboratory (20). Our verified novel gene predictions showed a significantly more restricted pattern of expression (Fig. 4*A*). The mean number of tissues for our positive predictions was 6.3, and 33% of the positive predictions showed expression in three or fewer tissues; the corresponding numbers for the mouse orthologs of human chromosome 21 genes are 8.2 tissues on average and 14% showing expression in three or fewer tissues. This difference in expression specificity was statistically significant (ANOVA, $F = 23.22$, df $= 1$, $P < 0.001$).

To determine whether prediction of pseudogenes by our method could explain some of the RT-PCR negatives, we computed the ratio of nonsynonymous to synonymous substitution rates (K_A/K_S) (21) for the subset of tested mouse predictions with unique putative human orthologs (Fig. 4*B*). The mean for PCR-positive predictions was 0.29 whereas for PCR-negative predictions it was 0.72. The difference was statistically significant (ANOVA, $F = 34.86$, df $= 1$, $P < 0.001$), suggesting that (*i*) some of the negative predictions may be pseudogenes, and (*ii*) K_A/K_S can be efficiently incorporated in the enrichment protocol to increase specificity (22).

Among the predictions with confirmed splices, 112 had significant homology to known genes and/or domains. A few of these genes, which were not represented in databases at the beginning of our gene survey, were submitted to databases and/or published in the literature in the intervening months. For example, we correctly predicted the first four protein coding exons of *TRPV3*, a heat-sensitive TRP channel in keratinocytes (23), and both exons of *RLN3* (*preprorelaxin 3*), an insulin-like prohormone (24). The verified predictions with the most notable homologies are shown in Table 2, including a novel homolog of dystrophin that is discussed in the mouse genome paper (1). Table 2 includes two noncanonical homeobox genes, one that is most similar to fruitfly brain-specific homeobox protein (Figs. 2 and 3 *A* and *B*) (25) and another that is a Not-class homeobox, likely to be involved in notochord development (26). Four predicted genes were found to be expressed in the brain and are likely to have neuronal functions, including one paralog each of: *Nna1*, which is expressed in regenerating motor neurons (27); an *N*-acetylated-α-linked-acidic dipeptidase, which hydrolyses the neuropeptide *N*-acetyl-aspartyl-glutamate to terminate its neurotransmitter activity (28); a novel γ-aminobutyric acid type B receptor, which regulates neurotransmitter release (29); and an Ent2-like nucleoside transporter, which modulates neurotransmission by altering adenosine concentrations (30). Other verified genes are likely to be important in muscle contraction (myosin light chain kinase homolog), degradation of cell cycle proteins (fizzy/CDC20 homolog), Wnt-dependent vertebrate development (Dapper/frodo homolog), and solute and steroid transport in the liver (solute transporter β). Homologs of two further genes predicted in our studies are associated with disease. *ATP10C*, an aminophospholipid translocase, is absent from Angelman syndrome patients with imprinting mutations (31), and *otoferlin,* which is mutated in a nonsyndromic form of deafness (32).

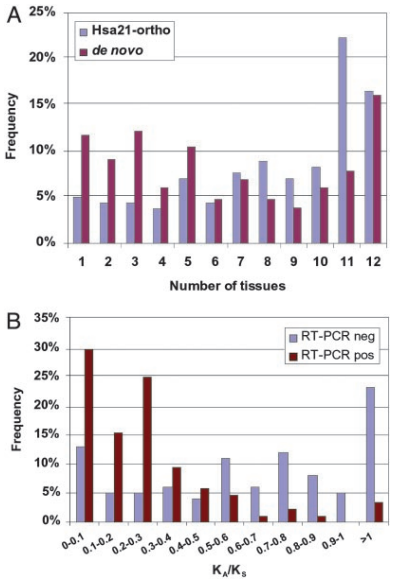


Fig. 4. Characteristics of verified predictions. (*A*) Expression specificity. Percentages of RT-PCR positive *de novo* predictions (red) and Hsa21 mouse orthologs (blue) expressed in 1–12 tissues, tested in the same cDNA pools. (*B*) Distributions of the ratio of nonsynonymous to synonymous substitution rate (K_A/K_S) in 83 RT-PCR positive (red) vs. 98 RT-PCR negative (blue) mouse predictions with reciprocal best BLAST matches among the human predictions.

GENETICS

Table 2. Novel mouse genes, their tissue expression, and their homologs

Code	B	H	K	Y	V	S	M	L	T	K	E	O	%Id	Ln	Homology
3B1								+	+				38	134	Dystrophin-like; with ZZ domain
3B3				+		+		+	+	+			25	184	Novel aquaporin; similar to *Drosophila* CG12251
3C3			+		+			+	+		+		25	260	TEP1 (telomerase associated); probable ATPase
3C5									+		+		47	198	Voltage-dependent calcium channel γ subunit
4B3			+			+			+				34	74	IFN-induced/fragilis transmembrane family
4C6		+				+		+	+	+			30	134	IL-22-binding protein CRF2-10
4G4	+								+	+	+		64	109	Nna1p, nuclear ATP/GTP-binding protein
5B5						+			+	+			43	111	Likely aminophospholipid flippase (transporting ATPase)
1E3	+			+	+			+			+		40	106	*N*-acetylated-α-linked-acidic dipeptidase (NAALADase)
6C4								+	+				42	117	Not-type homeobox; poss. involved in notochord development
6F5	+	+	+										66	102	*Drosophila* brain-specific homeobox protein (bsh)
11F2	+					+		+	+	+			29	216	Human γ-aminobutyric acid type B receptor 2, neurotransmitter release regulator
5A2			+		+	+			+				41	36	Skate liver organic solute transporter β
11B6				+				+			+		55	116	IFN-activatable protein 203; nuclear protein
12B3	+			+	+	+		+	+	+	+		25	229	Fatty acid desaturase; maintains membrane integrity
11F6	+	+				+		+	+	+	+		44	494	Rat vanilloid receptor type 1 like protein 1
12E3								+	+				52	175	Fizzy/CDC20; modulates degradation of cell-cycle proteins
12F1		+				+	+	+	+				43	355	Otoferlin (mutated in DFNB9, nonsyndromic deafness)
12H1	+	+							+				45	116	Fruitfly additional sex combs; a Polycomb group protein
12C4	+							+			+		43	133	*Caenorhabditis elegans* C15C8.2; single-minded-like; HLH and PAS domains
12D2						+							41	397	Cytosolic phospholipase A2, group IVB
12A5	+												38	415	Fruitfly GH15686p; Ent2-like nucleoside transporter
12E5	+			+				+				+	32	111	Relaxin 3 preproprotein; prohormone of the insulin family
11A1			+	+	+		+					+	89	75	Mouse BET3, involved in ER to Golgi transport
11A2	+	+						+	+		+	+	70	207	Vacuolar ATP synthase subunit S1
11B2							+	+	+	+	+	+	54	271	Myosin light chain kinase, skeletal muscle
11G2	+		+	+	+	+		+	+	+	+	+	36	179	Dapper/frodo (transduces Wnt signals by interacting with Dsh)

Code, Coding name of tested gene model. B, brain; H, heart; K, kidney; Y, thymus; V, liver; S, stomach; M, muscle; L, lung; T, testis; K, skin; E, eye; O, ovary. %Id, Percentage amino acid identity. Ln, Number of amino acids in the local alignment between the prediction and the homolog.

Discussion

We have demonstrated a remarkably efficient mammalian gene discovery system. This system exploits the draft mouse and human genome sequences in both an initial gene-prediction stage and an enrichment stage. The first stage consists of SGP2 and TWINSCAN, gene-prediction programs that use genome alignment in combination with statistical patterns in the DNA sequence. We have shown elsewhere that both programs have greater sensitivity and specificity than single-genome *de novo* predictors, such as GENSCAN (13, 14). In this article, we have demonstrated the effectiveness of the enrichment stage, in which predictions are retained only if the protein predicted in mouse aligns to a human protein predicted by the same program, with at least one predicted intron at the same location (aligned intron, Fig. 1). In our pool of predictions, the aligned intron filter is expected to eliminate 24 times more RT-PCR negatives than RT-PCR positives. This enrichment procedure can be applied to predictions from any program.

Our goal was to develop a low-cost, high-throughput system for finding and verifying coding regions that are missed by annotation systems that require existing transcript evidence. ENSEMBL was chosen as the representative of such systems because the Mouse Genome Sequencing Consortium judged it to be the most suitable tool for timely, cost-effective, reliable annotation of the mouse genome sequence. Thus, we evaluated our system by investigating genes that do not overlap ENSEMBL predictions. Our system is not designed to find genes that would be missed by expert manual annotators, who can effectively integrate information such as the predictions of GENSCAN (8) and GENOMESCAN (33), percent-identity plots (34), comparison to fish genomes (35, 36), alignment of weakly homologous proteins, and alignment of EST sequences. As a result, we did not exclude gene predictions from our evaluation based on these indicators.

Our two-stage system identified a highly reliable pool of 827 predicted genes not overlapping the standard annotation, of which we tested 154 for expression by using RT-PCR and direct sequencing. Primers designed for a single pair of adjacent exons in each predicted gene yielded a spliced PCR product whose sequence closely matched that of the predicted exons in 76% of these tests. In the only other published report of high-throughput verification of gene predictions of which we are aware, 14% of predictions not overlapping the standard annotation yielded spliced products (37). These numbers cannot be compared directly because of differences in the sampling criteria, but the magnitude of the difference suggests our method provides new levels of efficiency in experimental confirmation of genes outside the standard annotation set.

The sensitivity of our method also appears to be high. Predictions in our enriched pool overlap 90% of multiexon genes predicted by ENSEMBL. However, it has been estimated that >4,000 ENSEMBL predictions comprising 12,000 predicted exons are in fact pseudogenes (1). Although the precise number of multiexon pseudogenes in the ENSEMBL annotation is unknown, this estimate suggests that our enriched pool may overlap a much larger fraction of the functional genes identified by ENSEMBL. Further, RT-PCR tests of TWINSCAN and SGP2 predictions outside the enriched pool indicate that a relatively small number of these predictions are transcribed and spliced in the 12 tissues tested. Thus, the enrichment procedure is sensitive to both ENSEMBL predictions and verifiable predictions by TWINSCAN and SGP2.

Using our system, we confirmed one intron of 139 predicted genes that do not overlap any gene in the standard mouse genome annotation (1). Ninety-two of the RT-PCR positive introns (66%) did not align to any mouse EST, and these might have posed difficulties even for human annotators. Furthermore, seven of the RT-PCR negative introns (4%) did align to mouse ESTs and six of these were in the enriched pool, suggesting that the true percentage of transcribed and spliced predictions in this pool may be even higher than the RT-PCR positive percentage.

Among RT-PCR positive predictions, 24 had homologies to known proteins that we found particularly interesting (Table 2). The positive identification of these homologs is expected to impact numerous research programs devoted to genes of developmental and medical importance. In general, these genes were probably missed in the ENSEMBL annotation because the length and percent identity of the homologies were not sufficient to support a protein-based gene prediction (Table 2). In many cases, such as the predicted homolog of a brain-specific homeobox protein, the ex-

pression patterns we found were consistent with what would be expected from the function of the known homolog (Fig. 3 *A* and *B*).

The confirmed 139 genes also showed a relatively restricted expression pattern, on average. Because all mouse orthologs of genes on human chromosome 21 had already been tested by using the same experimental protocol and the same cDNA pools, we were able to directly compare expression patterns. To the extent that the known genes on chromosome 21 are no more tissue specific than the complete set of known genes, the results (Fig. 4) suggest that our system may be particularly sensitive to genes with tissue-restricted expression. Qualitatively similar restricted expression patterns were reported for novel GENSCAN predictions on chromosome 22 (37), lending further support to the value of *de novo* prediction for identifying genes with tissue-restricted expression.

Of the RT-PCR positive novel predictions, only 33% have identifiable homologs in the sequenced fish (*Fugu*/*Tetraodon*/zebrafish) genomes. Comparing this finding to the recent estimate that three-quarters of all human genes can be recognized in the *Fugu* genome (36) suggests that our system may be particularly sensitive to genes that are not ubiquitous in the vertebrate lineage. Genes with relatively restricted expression patterns and species distribution can be difficult to find by using transcript-based methods like GENEWISE (38) and compact-genome methods like EXOFISH (35), but they appear to be tractable for our system.

Extrapolating from the success rates in all categories, the expected total number of gene predictions that could be successfully RT-PCR amplified in the cDNA pools we tested is 1,019 (Table 1), adding ≈5% to the number of functional mouse genes identified by ENSEMBL (1). The number of distinct genes verifiable in this way may be slightly smaller, because the effect of fragmentation in ENSEMBL and in our predictions is not readily testable. However, the number of predictions that are transcribed and spliced is likely to be >1,019, because (*i*) we tested only one exon pair from each prediction and (*ii*) we used only 12 adult mouse tissues (20).

The relatively low success rate in the pools failing the enrichment step suggests that the number of real, multiexon genes whose existence has been predicted but not yet confirmed is in the range of 1,000–2,000 (including those predictions in the enriched pool that have not been confirmed). Because we have used only two prediction programs, TWINSCAN and SGP2, it is possible that other programs might yield a large additional set of predictions that pass the enrichment step. However, GENSCAN yields only 49 additional predictions that pass enrichment and novelty criteria and do not overlap the 1,428 "aligned intron" novel predictions from TWINSCAN and SGP2 (3%). These 49 are worth testing, and adding more prediction programs will yield at least a few more predictions with aligned introns. Nonetheless, the data presented here suggest that the 1,428 predictions in the enriched pool may overlap a significant fraction of the previously unannotated, multiexon mouse genes.

Using the draft sequences of the mouse and human genomes, we have developed a cost-effective, high-throughput system for predicting genes and verifying the existence of corresponding spliced transcripts. Applying this system to the entire mouse genome, we showed that an automated system can produce a large set of experimentally supported mammalian gene predictions outside the standard annotation. Further, the average cost per verified exon pair is less than two primer pairs and sequencing reactions. We expect that testing the remaining predictions in the enriched pool will locate most multiexon mouse genes that are currently unannotated, bringing us significantly closer to identification of the complete mammalian gene set.

As more mammalian genomes are sequenced, the need for experimentally validated high-throughput annotation will continue to grow, as will the data available for methods such as ours. Using the sequences of more genomes, it may be possible to extend this approach to single-exon and lineage-specific genes. In combination with methods like ENSEMBL and refinement by expert annotators, these developments may bring complete, experimentally supported genome annotation within reach.

We are grateful to the Mouse Genome Sequencing Consortium for providing the mouse genome sequence as well as support throughout the analysis process. We are particularly grateful to Eric Lander, Robert Waterston, Ewan Birney, Adam Felsenfeld, and Ross Hardison for advice and encouragement. Thanks are also due to Marc Vidal, Lior Pachter, Kerstin Lindblad-Toh, and Gwen Acton for participation in pilot experiments and Tamara Doering for helpful comments on the manuscript. Research at Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra/Centre de Regulació Genòmica is supported by a grant from the Spanish Plan Nacional de Investigacion y Desarrollo. J.F.A. is supported by a fellowship from the Instituto de Salud Carlos III. The Division of Medical Genetics is supported by the Swiss National Science Foundation, National Centres of Competence in Research Frontiers in Genetics, and the Childcare and J. Lejeune Foundations. Research at Washington University was supported by Grant DBI-0091270 from the National Science Foundation (to M.R.B.) and Grant HG02278 from the National Institutes of Health (to M.R.B.).

1. Mouse Genome Sequencing Consortium (2002) *Nature* **420,** 520–562.
2. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., *et al*. (2002) *Nucleic Acids Res.* **30,** 38–41.
3. Pruitt, K. D. & Maglott, D. R. (2001) *Nucleic Acids Res.* **29,** 137–140.
4. Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., *et al*. (2001) *Nature* **409,** 685–690.
5. The FANTOM Consortium and The RIKEN Genome Exploration Research Group Phase II Team (2002) *Nature* **420,** 563–571.
6. Bairoch, A. & Apweiler, R. (2000) *Nucleic Acids Res.* **28,** 45–48.
7. Gasteiger, E., Jung, E. & Bairoch, A. (2001) *Curr. Issues Mol. Biol.* **3,** 47–55.
8. Burge, C. & Karlin, S. (1997) *J. Mol. Biol.* **268,** 78–94.
9. Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T. & Guigó, R. (2001) *Genome Res.* **11,** 1574–1583.
10. Pachter, L., Alexandersson, M. & Cawley, S. (2002) *J. Comput. Biol.* **9,** 389–399.
11. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. (2000) *Genome Res.* **10,** 950–958.
12. Bafna, V. & Huson, D. H. (2000) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8,** 3–12.
13. Parra, G., Agarwal, P., Abril, J. F., Wiehe, T., Fickett, J. W. & Guigó, R. (2003) *Genome Res.* **13,** 108–117.
14. Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. (2003) *Genome Res.* **13,** 46–54.
15. Korf, I., Flicek, P., Duan, D. & Brent, M. R. (2001) *Bioinformatics* **17,** Suppl. 1., S140–S148.
16. Parra, G., Blanco, E. & Guigó, R. (2000) *Genome Res.* **10,** 511–515.
17. Guigó, R., Knudsen, S., Drake, N. & Smith, T. (1992) *J. Mol. Biol.* **226,** 141–157.
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
19. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
20. Reymond, A., Marigo, V., Yaylaoglu, M. B., Leoni, A., Ucla, C., Scamuffa, N., Caccioppoli, C., Dermitzakis, E. T., Lyle, R., Banfi, S., *et al*. (2002) *Nature* **420,** 582–586.
21. Hughes, A. L. & Nei, M. (1988) *Nature* **335,** 167–170.
22. Nekrutenko, A., Makova, K. D. & Li, W. H. (2002) *Genome Res.* **12,** 198–202.
23. Peier, A. M., Reeve, A. J., Andersson, D. A., Moqrich, A., Earley, T. J., Hergarden, A. C., Story, G. M., Colley, S., Hogenesch, J. B., McIntyre, P., *et al*. (2002) *Science* **296,** 2046–2049.
24. Bathgate, R. A., Samuel, C. S., Burazin, T. C., Layfield, S., Claasz, A. A., Reytomas, I. G., Dawson, N. F., Zhao, C., Bond, C., Summers, R. J., *et al*. (2002) *J. Biol. Chem.* **277,** 1148–1157.
25. Jones, B. & McGinnis, W. (1993) *Development (Cambridge, U.K.)* **117,** 793–806.
26. Talbot, W. S., Trevarrow, B., Halpern, M. E., Melby, A. E., Farr, G., Postlethwait, J. H., Jowett, T., Kimmel, C. B. & Kimelman, D. (1995) *Nature* **378,** 150–157.
27. Harris, A., Morgan, J. I., Pecot, M., Soumare, A., Osborne, A. & Soares, H. D. (2000) *Mol. Cell. Neurosci.* **16,** 578–596.
28. Pangalos, M. N., Neefs, J. M., Somers, M., Verhasselt, P., Bekkers, M., van der Helm, L., Fraiponts, E., Ashton, D. & Gordon, R. D. (1999) *J. Biol. Chem.* **274,** 8470–8483.
29. Billinton, A., Ige, A. O., Bolam, J. P., White, J. H., Marshall, F. H. & Emson, P. C. (2001) *Trends Neurosci.* **24,** 277–282.
30. Crawford, C. R., Patel, D. H., Naeve, C. & Belt, J. A. (1998) *J. Biol. Chem.* **273,** 5288–5293.
31. Meguro, M., Kashiwagi, A., Mitsuya, K., Nakao, M., Kondo, I., Saitoh, S. & Oshimura, M. (2001) *Nat. Genet.* **28,** 19–20.
32. Yasunaga, S., Grati, M., Cohen-Salmon, M., El-Amraoui, A., Mustapha, M., Salem, N., El-Zir, E., Loiselet, J. & Petit, C. (1999) *Nat. Genet.* **21,** 363–369.
33. Yeh, R. F., Lim, L. P. & Burge, C. B. (2001) *Genome Res.* **11,** 803–816.
34. Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. & Miller, W. (2000) *Genome Res.* **10,** 577–586.
35. Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., *et al*. (2000) *Nat. Genet.* **25,** 235–238.
36. Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al*. (2002) *Science* **297,** 1301–1310.
37. Das, M., Burge, C. B., Park, E., Colinas, J. & Pelletier, J. (2001) *Genomics* **77,** 71–78.
38. Birney, E. & Durbin, R. (2000) *Genome Res.* **10,** 547–548.
39. Notre dame, C., Higgins, D. G. & Heringa, J. (2000) *J. Mol. Biol.* **302,** 205–217.

Discussion

The rapid release of completed genome sequences has lead to significant developments in genome annotation and gene finding tools. In the near future, comparative approaches will be essential in order to build a more accurate catalog of genes for each organism. Previous chapters have described the development of different tools and protocols to obtain sets of gene predictions that outperform the currently available methods. In what follows, we discuss the utility of the *ab initio* gene finding methods versus the comparative approaches, the utility of the generated data sets and how gene prediction can lead high throughput experimental approaches. We also summarize some open problems in computational gene prediction and explore the future trends in the field.

6.1 geneid

The results previously presented indicate that the current version of `geneid` has an accuracy comparable to the most used gene prediction programs (mostly based on generalized Hidden Markov Models, GHMMs). In a GHMM approach, gene structure components are characterized with states and the gene model is generated according to the probabilities of transition and emission of the state machine. Therefore, the cost of finding the optimal parse in a GHMM is considered to be quadratic in the number of possible state transitions and linear to the length of the sequences (Burge and Karlin, 1997). In `geneid`, however, the simplicity of the algorithm architecture (signals $\rightarrow$ exons $\rightarrow$ genes) is reflected in a more efficient implementation that is asymptotically linear only in terms of length of the input sequences, without any loss in the accuracy of the results. The main difference between both approaches relies on the separation in `geneid` between signal and exon generation and the gene assembly stage. In the GHMM based gene finders, however, both tasks are performed simultaneously.

Most gene finders suffer from lack of specificity, predicting a large number of false positive exons and genes particularly in long genomic regions. Comparatively, `geneid` has superior specificity than other existing gene prediction programs, showing a more conservative behavior. The price is paid in terms of sensitivity. In general, for large genomic sequences encoding multiple genes, the overall accuracy of `geneid` is comparable, to that of the most accurate existing tools, but offers a better balance between specificity and sensitivity.

The parameter file is text based and it has a clear and direct interpretation. The statistical models for the recognition of the sequences signal are position weight arrays (PWAs). A Markov Model chain is used as coding statistic. Both are computed as log-likelihood ratios from real exons. These models are specially useful because the number of dependencies, and thus the number of parameters, can be adapted depending on the amount of available data. The order of the dependencies can be incorporated in the model without any modification in the source code.

These properties make `geneid` specially useful for the new genomes that are going to come. Although we are planning to produce parameter files for many species, the simplicity in the specification of the parameters allows any advanced user to generate his own specific parameter file. In the *Annexed papers* section there are two examples of the `geneid` training in different species. In these two cases, `geneid` was a key tool for the annotation of *Dictyostelium discoideum* (Glokner et al., 2002) and *Tetraodon nigroviridis* (Tetraodon Genome Sequencing Consortium, 2004) genomes.

The flexibility of `geneid` gene model allows the user to introduce any type of external information to be incorporated with the *ab initio* predictions. Recently this capability proved to be very effective in the prediction of selenoproteins by the introduction of predicted mRNA secondary structures.

In selenoproteins, the presence of a secondary structure (SECIS element) in the 3' UTR of the mRNA induces the UGA codon, usually a termination signal, to be translated as selenocysteine, the 21st amino acid. In consequence, current computational gene prediction methods, which rely on the standard meaning of sequence signals, invariably mispredict selenoprotein genes. In order to address this problem, a slight modification in `geneid` was introduced to permit the dual meaning of the UGA triplet. However, the inclusion of an additional exon-defining signal decreases the overall gene prediction accuracy. Such a drawback can be rapidly compensated with the introduction of *ad hoc* biological knowledge into the `geneid` prediction to constrain which genes may be interrupted by in-frame TGA codons. In this case, the downstream position of potential SECIS elements along the genome is informative, and delimits where selenoprotein genes can lie. This strategy, together with comparative approaches, has been successfully applied to describe the *D. melanogaster*, *H. sapiens* and *T. rubripes* selenoproteomes (Castellano et al., 2001; Kryukov et al., 2003; Castellano et al., 2004).

Among the drawbacks of `geneid` is the somehow less rigorous probabilistic treatment of the scoring schema. For instance, the exon weight (*EW*) value that could be considered as the prior odds of being a real exon versus not being an exon must be computed for each parameter file. Another limitation of `geneid` is the lack of a model of exon and intron length distributions. The current version of `geneid` also lacks a model of the distributions of the number of exons per gene. We are currently working on some modification of the algorithm to allow the definition of intron and exon length distribution and to model the variability of the number of exons for each gene.

The flexibility of `geneid` also facilitates its ongoing development. Our short-term plans include the definition of a branch point model, the modeling of long distance dependencies in splice sites, the prediction of UTRs and the analysis of suboptimal gene structures to determine different alternative splicing isoforms.

6.2 sgp2

We have described a successful way to combine `geneid` and information from comparative genomic data into `sgp2`. As expected, obtained results indicate that, by including information from genome sequence conservation (for instance, using the human and mouse genomes), `sgp2` clearly outperforms *ab initio* gene finding approaches.

Our approach, based on local alignments, allows to use a draft genome or shotgun sequences. This means that as the sequence of a new genome starts becoming available (shotgun sequences), it can be used to improve the gene predictions of other preexisting genomes. Another important feature is the ability of `sgp2` to make the predictions on top of a set of annotated genes. This has been proved to be specially useful in genome-wide annotations in order to reduce the number of joint and split predicted genes.

In comparison to existing algorithms, `sgp2` is most similar in its overall architecture to the recently developed `twinscan` (Korf et al., 2001). `twinscan` uses `blastn` alignments and different nucleotide conservation models integrated in a GHMM to generate the gene predictions. `sgp2`, instead, use `tblastx` alignments obtained with the corresponding amino acid substitution matrix. This will probably help `sgp2` to find similarity between sequences from distant species were the conservation at nucleotide level could be more difficult to find. On the other hand, `twinscan` and the model of conservation at nucleotide level will be more useful for the recognition of more closely related species. In close related species, coding sequences will be more conserved and little changes in the pattern of mutation, not visible at amino acid level, can be captured by the nucleotide conservation models used by `twinscan`.

`sgp2` is flexible enough so that it can be easily accommodate to analyze species different pairs of genomes than human and mouse. So far, `sgp2` is currently used in the prediction of the genes in *G. gallus* using *H. sapiens* as a reference genome.

In addition to this, we are starting to analyze the possibility of combining alignments obtained from the comparison of more than one species at the same time. However, to achieve this objective we need to analyze in depth the effect of the substitution matrix used for species at different evolutionary distances.

One of the limitations of any comparative gene prediction method is that the target sequence must have appropriate informant sequences, with `sgp2` ability to utilize unfinished informant sequences and with the current rate of genome sequencing, this will become a minor restriction.

Further plans in the development of `sgp2` include the measurement of the synonymous versus non-synonymous substitution rates in the alignments and conservation of the splice signals in the informant genome.

6.3 Ab initio vs. comparative gene prediction

The analysis of the performance of gene finding methods has shown that comparative methods clearly outperform standard *ab initio* approaches. Although, sensitivity in comparative gene finders does not clearly improve, the specificity is superior. Therefore, one

can ask the question of whether there is any reason to keep developing such *ab initio* gene finding methods.

From our point of view, the question has a positive answer. First of all, because the core of all comparative gene finding programs relies, at least partially, on *ab initio* gene recognition methods. Therefore, improvements in *ab initio* recognition models can then be applied to comparative approaches.

Another important reason for developing *ab initio* gene prediction tools, is that, although we have a lot of complete genomic sequences, is not always possible to find a reference genome of a species at the correct evolutionary distance to use comparative gene prediction. However, as mentioned in the previous section, with the current rate of genome sequencing, this will become a minor restriction.

Finally, *ab initio* gene prediction programs should be considered as applications, which make it possible to gather our knowledge about what a gene is. This elucidates that our current knowledge of the genome biology is rather poor: we are not able to reproduce what the cell does to obtain the proteins encoded in a genome. Even using sophisticated probabilistic models we still predict less than 50% of the human genes correctly. Probabilistic models only make sense when the underlying model is biologically meaningful, and the training samples are not biased. Therefore, we believe that future *ab initio* developments will tend to be more based on biological information and less on pure statistical data.

New models should try to mimic the biological scenario, and the actual underlying processes. Then, apart from the specific goal of gene prediction, new approaches will probably be able to determine effects of point mutations or single nucleotide polymorphisms in the pattern of transcription and the alternative splicing events.

6.4 Evolution of the signals that define genes

The fact that individual parameter files for each species perform better than general models can be explained because each genome seems to have its own signatures for gene signaling which were shaped by evolutionary pressures. This implies that gene structures and the transcription and translation machinery in each species are adapted, enabling the cell to appropriate transcription and translation for each genome.

During the process of building different parameter files for `geneid`, a complete database of curated genes for different species has been generated. The study of the differences on the specification of genes can elucidate the evolution of the mechanisms involved in gene expression and help us to better understand the underlying processes.

In this regard, a recent publication by Korf (2004) started to analyze the results of predictions generated in one species using parameters for many different species. Korf (2004) shows that the most compatible parameters may not come from the nearest phylogenetic neighbor. He trained and evaluated `snap`, a gene finding program based on `genscan`, in the *Arabidopsis thaliana*, *Ceanorhabditis elegans*, *Drosophila melanogaster*, and *Oryza sativa* genomes, and demonstrate that for instance in *A. thaliana*, the best foreign parameters come from *C. elegans*instead of *Oryza sativa* (rice).

Some of these specific variations seem to be related with the general C+G content of

the genome. Other features, like splicing, seem to have different recognition mechanism (intron versus exon definition) that should have different models for each species. For instance most fungi seem to have a very conserved branch point, while some plants seem to have an intron signaling system through differential intron base content. Therefore, gene prediction programs should adapt their model to the specific prevalences and nature of each signaling mechanism. There might not be a single model of how genes are defined. Rather, different mechanisms, that could have evolved independently, are acting together in each species to recognize and process its genes.

In addition, analysis of gene signatures may also help to resolve conflicting phylogenies and to pinpoint horizontal gene transfers, genetic drifts and other evolutionary events. If we are able to recognize different gene signatures, we could easily identify genes that have been recently incorporated into a genome, and track down where they come from.

6.5 Conservation of the exonic structure

The availability of an increasing number of eukaryotic genomes is contributing to the understanding of the evolution of exonic structure. Comparative analysis of exonic structure and splice signals of homologous genes from different species will certainly contribute to our understanding of the mechanisms by which splice sites are recognized.

Recent large scale comparative analyses have reported extraordinary conservation of the exonic structure between human and mouse orthologous genes (Roy, 2003). The exonic structure conservation has been successfully used in the filtering protocol that we have shown to target *bona fide* genes among thousands of computational gene predictions.

Some other genomic approaches have been used to exploit exon structure conservation. In one of such approaches, Dewey et al. (2004) describes a method for the simultaneous prediction of homologous genes with identical structure in the human, mouse and rat genomes using `slam`. `slam` is a gene prediction program based on pair-GHMMs, where alignment and gene prediction are performed simultaneously. The combination of pairwise predictions made with `slam` provides 3698 gene triplets in the human, mouse, and rat genomes which are predicted with exactly the same gene structure. These consensus predicted genes greatly improve the specificity (over 90% of the predicted structures correspond to complete actual genes), but at expenses of a large loss of sensitivity.

Although these approaches based on pair-GHMMs or phylo-GHMMs (where more than two species are treated simultaneously) are very promising, they have not yet yielded practical improvements in the accuracy of gene prediction. Some limitations of these methods are that they need accurate complete syntenic maps between species and that the complexity of the evolutionary models leaves little room for complex gene structure models.

Nevertheless, gene prediction programs that exploit alignment and exonic structure conservation among multiple species are likely to outperform current gene finding methods in the coming years.

6.6 Experimental validation of the predictions

The emergence of high throughput techniques, characteristic of genomics research, has lead to the so-called data- or discovery- driven biology, in which data is obtained without the need for a hypothesis about the nature of any biological problem, as opposed to the classical hypothesis-driven approach in which experiments are performed (and data obtained) to test previously formulated hypothesis within the framework of a pre-existing theory.

Genome projects are mostly high throughput biology, and they certainly produce a lot of valuable data. High throughput biology alone, however (through indiscriminate sequencing of cDNA libraries), appears to have reached a limit in its ability to annotate genes in the human genome. For instance, we now start to see regions of the genome that are transcribed but do not appear to be coding for proteins. It is therefore time for the computational biologist to generate gene models with the enough confidence to be worth trying to be validated with high throughput experimental approaches.

Synergy between computational and experimental methods of gene identification will facilitate the full analysis of the currently sequenced genomes. As more genomes are sequenced, the need for experimentally validated high throughput annotation will continue to grow, as will the data available for such methods.

In this regard, recent reports underscore the importance of a hybrid approach. In one such reports, Tenney et al. (2004) shows that this approach could be feasible, at least by now, in genomes like *Cryptococcus neomorfans*. They argue that *C. neomorfans* is an attractive system for RT-PCR based annotation because it has relatively complex gene structures, while being a single-celled organism. This simplifies the task of obtaining representative mRNA samples. Application of computational gene prediction followed by experimental verification by RT-PCR has lead to the identification 63 complete novel genes. Now they are planning to extend this approach to validate the entire genome predictions.

Confirmation by RT-PCR and direct sequencing seems to be a cost effective technique that will probably constitute the basis for the final curated annotation of many available genomes. This approach is complementary to EST sequencing which produces data from highly expressed genes at a lower cost. However, RT-PCR of predictions is much more sensitive for genes that are expressed at relatively low levels and therefore, more difficult to obtain through ESTs sequencing. The success of these studies, suggests a new paradigm in high throughput genome annotation, in which gene predictions serve as the hypothesis that drives experimental determination of intron-exon structures.

6.7 Gene finding: open problems

Existing gene finding programs, although significantly advanced over those that were available a few years ago, still have several important limitations. Almost without exception, computational gene finders predict only the coding fraction of a single spliced form of non-overlapping, canonical protein-coding genes. Some key problems and future challenges in the gene prediction field are:

- To identify the untranslated regions of genes.
- To predict alternative transcripts. Alternative splicing events will be one of the most important problems to solve in the near future.
- To have a better characterization of the splicing enhancers and silencers that mediate alternative splicing, to allow models to predict alternative exons or aberrant splicing events.
- To improve our understanding of CpG islands, methylation patterns and G+C variations across the genome, and to use this information to improve gene predictions.
- To identify gene promoter regions and the corresponding transcription start site.
- To characterize promoter regions, to be able to elucidate the combination of transcription factors needed for the activation and inhibition as well as the tissue and developmental stage specific expression pattern.
- To predict genes that encode for functional RNAs.
- To predict insulators, matrix-attachment regions and nucleosome organization patterns that could play a key role in the accessibility of the transcriptional machinery to the chromatin.
- To predict uncommon features as overlapping genes, genes within introns, genes with non canonical splice sites, mRNA editing or frame shifting. We assume these cases to be rare, but because these assumptions are implicit in our gene models, we may have been seriously underestimating their occurrence.

At the root of these limitations lies our still incomplete knowledge of what defines an eukaryotic gene, and which mechanisms are mediating the recognition of sequence signals involved in gene identification and processing in the eukaryotic cell. The models in which current gene finding methods are based are over-simplistic and only include a partial knowledge of gene biology. In most cases computational programs detect genes mostly by the imprinting they leave on the sequence, like coding statistics, that can be considered the consequence, but not the cause of their existence.

We believe that the problem should be addressed in a more restricted scenario. Instead of trying to predict genes in complete genomic sequences, try to divide the general problem of gene finding into easier and biologically meaningful sub-problems.

For instance, promoter regions and the transcription factor binding sites that define them seem to be recognized as a combination of not very conserved motifs. Using current pattern searching tools, putative binding sites are predicted all over the genomic sequence. Therefore, promoter and the transcription start site of each transcript may be intrinsically related to the structure of the DNA in the nucleus, the attachment of the chromatin, the nucleosome organization and the methylation patterns. Improvements in the knowledge of the accessibility of the transcription to the DNA will be crucial to solve this puzzle. Recently, Bajic and Seah (2003) improved the prediction of trancriptional start sites using information of CpG islands and signals in the downstream promoter region. Prediction of promoter regions and transcriptional start sites could be considered a field on its own.

Modeling of splicing is another important open problem. Even the most sophisticated computational models of splicing currently available are limited to model dependencies between positions within the canonical signals defining the intron boundaries. The models implicitly assume, in consequence, the splice signals to be recognized independently and atemporally in a nucleic acid sequence without further information.

There is, however, increasing experimental evidence suggesting that additional intronic and exonic sequences play a role in the definition of the intron boundaries, and in the regulation of the production of alternative splice forms. Moreover, it must be taken into account that transcription and splicing seem to occur at the same time in what is called the "mRNA factory" (Zorio and Bentley, 2004). There are dynamic relations, not yet completely understood, between transcription and splicing, and dependencies between distant splice signals can not be discarded. Thus, while the gene is transcribed there may be some pattern of splice site recognition depending on the speed, length, nucleotide composition and the accessibility of the splice sites among other possibly relevant features. RNA structure, too, may influence splice site selection.

All these phenomena should be taken into account in a biologically realistic model of the splicing process. The results of future gene prediction tools should not be an unique model for each gene, but a set of putative spliced transcripts with the associated expected frequency. Additionally, while experimental data is crucial to understand the mechanistic details of these phenomena, the fact that we have accumulated in our databases a large collection of annotated splicing events, makes the contribution of computational analyses very important.

In summary, only with enough biological information of the underlying mechanisms, gene prediction will be transformed from being statistical to being biological in nature. However, computational analyses will be decisive to direct the efforts invested to get this biological knowledge and providing hypotheses to be experimentally tested.

Conclusions

The following conclusions can be drawn from this dissertation:

1. The results presented here, indicate that the current version of `geneid` shows an accuracy comparable, and often superior, to the most currently used methods. In favour of `geneid` is the simplicity and modularity of its structure.

2. Gene recognition patterns seem to be conserved over large phylogenetic distances, but they also appear to have some taxa specific components. For instance, although the canonical consensus splicing sequences are conserved (GT-AG), the pattern of conservation around these sites among species differs notably. This variability suggests specific adaptations of the cell machinery to the recognition and processing of genes.

3. Since the signals that define genes seem to have some species-specific signatures, parameters and models for each species improve the prediction of genes. The construction of `geneid` parameter files for different species showed an important improvement in the accuracy of the predictions.

4. Our experiments demonstrate that integrating *ab initio* information with genomic similarity, even from shotgun reads, using `sgp2`, significantly improves accuracy over *ab initio* standard methods.

5. The enrichment prediction protocol, based on exonic structure conservation between closely related species, has led to an increase of the amplification success ratio of predicted genes from 3% to 76%. This experiment has proved the value of comparative genomics and the conservation of the gene structure in gene finding.

6. The synergy between computational and experimental methods of gene identification has shown to yield hundreds of novel human genes. The success of our study, suggests a new paradigm in high throughput genome annotation, in which gene predictions serve as the hypothesis that drives experimental determination.

Annexed Papers

In this section are gathered the other relevant papers I have collaborated in. In these cases my participation was less relevant than in the ones showed in the main block. Before each article there is a little description of my contribution to each work.

Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*

G. Glökner, L. Eichinger, K. Szafranski, J.A. Pachebat, A.T. Bankier, P.H. Dear, R. Lehmann, C. Baumgart, G. Parra, J.F. Abril, R. Guigó, K. Kumpf, B. Tunggal, the Dictyostelium Genome Sequencing Consortium, E. Cox, M.A. Quail, M. Platzer, A. Rosenthal and A.A. Noegel.
Nature 418(6893):79-85 (2002)

Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotipe

Tetraodon Genome Sequencing Consortium (including G. Parra and R. Guigó).
Nature (431):946-957 (2004)

Initial sequencing and comparative analysis of the mouse genome

Mouse Genome Sequencing Consortium (including G. Parra and R. Guigó).
Nature 420(6915):520-562 (2002).

Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*

G. Glökner, L. Eichinger, K. Szafranski, J.A. Pachebat, A.T. Bankier, P.H. Dear, R. Lehmann, C. Baumgart, G. Parra, J.F. Abril, R. Guigó, K. Kumpf, B. Tunggal, the Dictyostelium Genome Sequencing Consortium, E. Cox, M.A. Quail, M. Platzer, A. Rosenthal and A.A. Noegel.
Nature 418(6893):79-85 (2002)

This work was done in collaboration with the Dictyostelium Genome Sequencing Consortium, which is an international consortium for the sequencing and the analysis of the genome of *D. discoideum*. The Dictyostelium Genome Sequencing Consortium is a collaboration between the University of Cologne, the Institute of Molecular Biotechnology in Jena, the Baylor College of Medicine in Houston, Institut Pasteur in Paris, and the Sanger Center in Hinxton.

D. discoideum is a soil-living amoeba with a very peculiar cell cycle: it grows as separate, independent cells but interact to form multicellular structures when challenged by adverse conditions such as starvation. Thousands of individual cells signal each other by releasing the chemo-attractant cAMP and aggregate together by chemo-taxis to form a multicellular structure that is surrounded by an extracellular matrix. This organism has unique advantages for studying fundamental cellular processes with powerful molecular genetic tools. These processes include chemo-taxis and signal transduction, and aspects of development such as cell pattern formation and cell-type determination. Many of these cellular behaviors and biochemical mechanisms are either absent or less accessible in other model organisms.

The hereditary information of *Dictyostelium* is contained in six chromosomes with sizes ranging from 4 to 7 Mb resulting in a total of about 34 Mb of haploid DNA genome with a base composition of 77% of adenines and thymines. This extreme base composition biased to A+T nucleotides have some influence in the signals and the codon usage that are used to codify the genes. Obviously, with such a biased base composition the accuracy of available gene prediction programs was very low.

Building a parameter file

A parameter file for `geneid` based on experimental annotated sequences from *D. discoideum* was generated. The training set was obtained by screening GenBank database (release 120.0, October 2000) for entries containing: "*Dictyostelium discoideum*" [organism] AND "complete" [title word] AND "CDS" [title word]. From the previous search, 160 sequences were obtained. 16 entries corresponding to mithocondrial or plasmid DNA were discarded. The quality of the sequences was checked following the criteria described in section 3.3.2. After the filtering protocol , 7 cases were discarded because the CDS was incomplete, 5 because of non-standard splice sites and 2 because of stop codons in frame.

Finally, 130 genomic sequences were gathered, from which 97 corresponded to multiexon and 33 to single-exon gene. An extra set of 250 mRNAs, provided by the Dictyostelium Genome Sequencing Consortium, was included in the training set. Thus, the

	Base level			Exon level				
	Sn	Sp	CC	Sne	Spe	SnSp	ME	WE
H. sapiens `geneid`	0.87	0.98	0.89	0.24	0.37	0.31	0.39	0.06
D. melanogaster `geneid`	0.12	0.83	0.22	0.03	0.08	0.06	0.66	0.23
A. thaliana `genscan`	0.79	0.98	0.83	0.29	0.36	0.33	0.32	0.08
P. falciparum `glimmer`	0.91	0.97	0.91	0.32	0.41	0.36	0.25	0.07
D. discoideum `geneid`	0.99	0.97	0.97	0.76	0.75	0.76	0.06	0.06

Table 7.1: Gene structure predictions statistics. Sn sensitivity, Sp specificity, ME missing exons and WE wrong exons, CC and SnSp correlations between sensitivity and specificity.

final set contained 380 translational starts sites, 170 pairs of splice sites and 472,549 coding nucleotides (96,542 bp from the 130 genes and 376,007 bp from the mRNA set).

PWAs of order zero (equivalent to PWMs) were computed for the start sites and splice sites. The corresponding informative positions were twelve bases upstream of the translation start site (methionine) and six positions downstream the coding region. For the donor and acceptor nucleotides in the region -2 to 6 and -15 to 1 bases were taken respectively (with 1 being position after the cleavage of the splice site, see Figure 3.3).

A Markov chain of order 5 was computed using the coding regions of the 130 sequences plus the 250 coding regions of the mRNA sequences, as a background model the intronic regions were used.

After the exhaustive search process on the 130 sequences an *EW* of -9.50 was found to optimize the value of the correlation coefficient at nucleotide level.

Prediction accuracy on Dictyostelium

To asses the accuracy of the available gene prediction methods in *D. discoideum*, different programs were selected. `glimmer` (Delcher et al., 1999) was selected, because it had a version with parameters for *Plasmodium falciparum*, which genome has a similar G+C content and similar gene structures. `glimmer` uses interpolated Markov models to find genes in microbial DNA. The new release based on *Plasmodium falciparum* allows multigenic and multiexonic predictions. `genscan` was also selected and it was used with *Arabidopsis thaliana* parameter file. *A. thaliana* genes have some similarities with the *D. discoideum* gene structure. Introns are short and with a low G+C content and long stretches of adenines and thymines. `geneid` with human and *Drosophila* parameters were also selected.

Table 7.1 shows that `geneid` using *D. melanogaster* parameter file has very low accuracy. A lot of real exons were lost (with a missing exon rate of 0.66) and a lot of wrong exons were generated (with a wrong exons rate of 0.23). Low accuracy at base level and at exon level could be explained because neither the composition nor the signals of *D. melanogaster* seems to be similar to *D. discoideum*(See Figure 3.3 and Table 3.2). `genscan` using *A. thaliana* parameter file and `geneid` using human parameter file had similar performance. Although these species are not close to *D. discoideum*, both have regions of very low C+G content, and both programs are prepared to deal with low C+G content regions.

Sensitivity at base level was very high (0.79 using *A. thaliana* parameters and 0.87 using human parameters). However, the sensitivity at exon level was low (near 0.30). This low exon prediction accuracy could be explained by a different gene structure and splice sites signaling between those species. From the available tools `glimmer` (trained on *P. falciparum* sequences) had the most accurate predictions. As it was expected due their similarity in gene structure (a low number of short introns), and genomic C+G content. The correlation coefficient at base level was very high (0.91). However, at exon level the average between sensitivity and specificity was only of 0.36, revealing that the recognition of the splice signals was very poor. This is an important result because, although, these two species have important similarities at genomic level they seem to have differences in the splice sites definition. `geneid` trained on *D. discoideum* seemed to achieve the best results of all the programs, at the base and exon level, but we have to take in account that it could be over-training problems as far as the test and the training sets were the same.

Prediction of genes in chromosome 2 of *Dictyostelium discoideum*

`geneid` was run on all the contigs that correspond to the chromosome 2 of *D. discoideum*. Partial predictions or predictions shorter than 100 amino acid were discarded. The final annotation was based on the 2,799 genes predicted by `geneid`. The presented paper is focused on the analysis of the function and the structure of `geneid` predicted genes. This analysis reinforce the view that the evolutionary position of *D. discoideum* is located before the branching of metazoa and fungi but before the divergence of the plant kingdom.

sex-specific lethality that would accompany inappropriate somatic expression of *Sxl* (ref. 28). Moreover, using a very sensitive test[29], we determined that infection does not alter the effectiveness of the primary sex-determination signal (data not shown), perturbations of which can cause sex-specific lethality owing to inappropriate somatic expression of *Sxl*. Nevertheless, the possibility should be explored that *Wolbachia*-induced male killing reported for other *Drosophila* species[6] may be caused by inappropriate activation of *Sxl*.

Although it may seem surprising that infection with a parasite would reverse the deleterious effect of a mutation in the host genome, particularly when the isolation of that mutation had nothing to do with infection, such surprise should be tempered by the fact that the interaction described here between host and parasite mimics a naturally occurring situation mentioned above that was reported recently for the parasitic wasp *Asobara tabida*[9]. Moreover, in light of the fact that *Wolbachia* is a parasite that is known to manipulate host reproductive and sex-determination systems, it does not seem unreasonable that the host gene with which it interacts in *Drosophila* is the master regulator of sex-determination and a gene essential for oogenesis. The fact that the interacting gene in this case has been studied so extensively and belongs to a model experimental organism can be exploited to yield further insights into the mechanism by which this parasite takes advantage of its various arthropod hosts. □

Received 7 February; accepted 23 April 2002; doi:10.1038/nature00843.

1. Werren, J. H. Biology of *Wolbachia*. *Annu. Rev. Entomol.* **42,** 587–609 (1997).
2. Knight, J. Meet the Herod bug. *Nature* **412,** 12–14 (2001).
3. Werren, J. H. & O'Neill, S. L. in *Influential Passengers: Inherited Microorganisms and Arthropod Reproduction* (eds O'Neill, S. L., Hoffman, A. A. & Werren, J. H.) 1–41 (Oxford Univ. Press, Oxford, 1997).
4. Huigens, M. E. *et al.* Infectious parthenogenesis. *Nature* **405,** 178–179 (2000).
5. Bouchon, D., Rigaud, T. & Juchault, P. Evidence for widespread *Wolbachia* infection in isopod crustaceans: molecular identification and host feminization. *Proc. R. Soc. Lond. B* **265,** 1081–1090 (1998).
6. Hurst, G. D. D., Johnson, A. P., Schulenburg, J. H. G. & Fuyama, Y. Male-killing *Wolbachia* in *Drosophila*: a temperature-sensitive trait with a threshold bacterial density. *Genetics* **156,** 699–709 (2000).
7. Boyle, L., O'Neill, S. L., Robertson, H. M. & Karr, T. L. Interspecific and intraspecific horizontal transfer of *Wolbachia* in *Drosophila*. *Science* **260,** 1796–1799 (1993).
8. Bordenstein, S. R., O'Hara, F. P. & Werren, J. H. *Wolbachia*-induced incompatibility precedes other hybrid incompatibilities in *Nasonia*. *Nature* **409,** 707–710 (2001).
9. Dedeine, F. *et al.* Removing symbiotic *Wolbachia* bacteria specifically inhibits oogenesis in a parasitic wasp. *Proc. Natl Acad. Sci. USA* **98,** 6247–6252 (2001).
10. Bourtzis, K., Nirgianaki, A., Markakis, G. & Savakis, C. *Wolbachia* infection and cytoplasmic incompatibility in *Drosophila* species. *Genetics* **144,** 1063–1073 (1996).
11. Min, K. T. & Benzer, S. *Wolbachia*, normally a symbiont of *Drosophila*, can be virulent, causing degeneration and early death. *Proc. Natl Acad. Sci. USA* **94,** 10792–10796 (1997).
12. Cline, T. W. & Meyer, B. J. Vive la différence: males vs females in flies vs worms. *Annu. Rev. Genet.* **30,** 637–702 (1996).
13. Schupbach, T. Normal female germ cell differentiation requires the female X-chromosome to autosome ratio and expression of *Sex-lethal* in *Drosophila melanogaster*. *Genetics* **109,** 529–548 (1985).
14. Cook, K. R. *Regulation of Recombination and Oogenesis by the* ovarian tumor, Sex-lethal, *and* ovo *Genes of* Drosophila melanogaster. Thesis no. 381, Univ. Iowa (1993).
15. Salz, H. K., Cline, T. W. & Schedl, P. Functional changes associated with structural alterations induced by mobilization of a *P* element inserted in the *Sex-lethal* gene of *Drosophila*. *Genetics* **117,** 221–231 (1987).
16. Perrimon, N., Mohler, D., Engstrom, L. & Mahowald, A. P. X-linked female-sterile loci in *Drosophila melanogaster*. *Genetics* **113,** 695–712 (1986).
17. Bopp, D., Horabin, J. I., Lersch, R. A., Cline, T. W. & Schedl, P. Expression of the *Sex-lethal* gene is controlled at multiple levels during *Drosophila* oogenesis. *Development* **118,** 797–812 (1993).
18. Dines, J. L. *New Aspects of Functional Complexity for the Master Regulator of* Drosophila melanogaster *Sex Determination* Thesis no. 319, Univ. California, Berkeley (2001).
19. O'Neill, S. L., Giordano, R., Colbert, A. M. E., Karr, T. L. & Robertson, H. M. 16S rRNA phylogenetic analysis of the bacterial endosymbionts associated with cytoplasmic incompatibility in insects. *Proc. Natl Acad. Sci. USA* **89,** 2699–2702 (1999).
20. Bopp, D., Schutt, C., Puro, J., Huang, H. & Nothiger, R. Recombination and disjunction in female germ cells of *Drosophila* depend on the germline activity of the gene *Sex-lethal*. *Development* **126,** 5785–5794 (1999).
21. Dines, J., Lersch, B., Lu, B., Bell, M. & Cline, T. W. Functional specialization of SEX-LETHAL protein isoforms. *Annu. Drosophila Res. Conf. Program Abs. Vol.* **39,** a245 (1998).
22. Salz, H. K. *et al.* The *Drosophila* female-specific sex-determination gene, *Sex-lethal*, has stage-, tissue-, and sex-specific RNAs suggesting multiple modes of regulation. *Genes Dev.* **3,** 708–719 (1989).
23. Oliver, B., Perrimon, N. & Mahowald, A. P. Genetic evidence that the *sans fille* locus is involved in *Drosophila* sex determination. *Genetics* **120,** 159–172 (1988).
24. Steinmann-Zwicky, M. Sex determination in *Drosophila*: the X-chromosomal gene *liz* is required for *Sxl* activity. *EMBO J.* **7,** 3889–3898 (1988).
25. Pauli, D., Oliver, B. & Mahowald, A. P. The role of the *ovarian tumor* locus in *Drosophila melanogaster* germline sex determination. *Development* **119,** 123–134 (1993).
26. Page, S. L., McKim, K. S., Deneen, B., Van Hook, T. L. & Hawley, S. R. Genetic studies of *mei-P26* reveal a link between the processes that control germ cell proliferation in both sexes and those that control meiotic exchange in *Drosophila*. *Genetics* **155,** 1757–1772 (2000).
27. Hager, J. H. & Cline, T. W. Induction of female *Sex-lethal* RNA splicing in male germ cells: implications for *Drosophila* germline sex determination. *Development* **124,** 5033–5048 (1997).
28. Cline, T. W. A male-specific lethal mutation in *Drosophila melanogaster* that transforms sex. *Dev. Biol.* **72,** 266–275 (1979).
29. Cline, T. W. Evidence that *sisterless-a* and *sisterless-b* are two of several discrete 'numerator elements' of the *X/A* sex determination signal in *Drosophila* that switch *Sxl* between two alternative stable expression states. *Genetics* **119,** 829–862 (1988).

Acknowledgements

We thank L. Sefton for generating the original suppressed Sxl^{f4} strain, D. Presgraves for the *y w* CS *Wolbachia* strain, and B. J. Meyer for comments on the manuscript.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.W.C. (e-mail: sxlcline@uclink.berkeley.edu).

Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*

Gernot Glöckner*, Ludwig Eichinger†, Karol Szafranski*, Justin A. Pachebat‡, Alan T. Bankier‡, Paul H. Dear‡, Rüdiger Lehmann*, Cornelia Baumgart*, Genis Parra§, Josep F. Abril§, Roderic Guigó§, Kai Kumpf*, Budi Tunggal†, the *Dictyostelium* Genome Sequencing Consortium||Edward Cox¶, Michael A. Quail#, Matthias Platzer*, André Rosenthal||☆ & Angelika A. Noegel†

* *IMB Jena, Department of Genome Analysis, Beutenbergstr. 11, 07745 Jena, Germany*
† *Center for Biochemistry, Medical Faculty, University of Cologne, Joseph-Stelzmann-Str. 52, 50931 Köln, Germany*
‡ *Medical Research Council Laboratory of Molecular Biology, MRC Centre, Hills Road, Cambridge CB2 2QH, UK*
§ *Grup de Recerca en Informatica Biomedica, Institut Municipal d'Inveastigació Mèdica, Universitat Pompeu Fabra, Centre de Regulació Genòmica, 08003 Barcelona, Spain*
¶ *Princeton University, Princeton, New Jersey 08544, USA*
# *The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK*
☆ *Friedrich Schiller Universität, 07743 Jena, Germany*
|| *A full list of authors appears at the end of this paper*

The genome of the lower eukaryote *Dictyostelium discoideum* comprises six chromosomes. Here we report the sequence of the largest, chromosome 2, which at 8 megabases (Mb) represents about 25% of the genome. Despite an A + T content of nearly 80%, the chromosome codes for 2,799 predicted protein coding genes and 73 transfer RNA genes. This gene density, about 1 gene per 2.6 kilobases (kb), is surpassed only by *Saccharomyces cerevisiae* (one per 2 kb) and is similar to that of *Schizosaccharomyces pombe* (one per 2.5 kb)[1,2]. If we assume that the other chromosomes have a similar gene density, we can expect around 11,000 genes in the *D. discoideum* genome. A significant number of the genes show higher similarities to genes of vertebrates than to those of other fully sequenced eukaryotes[1–6]. This analysis strengthens the view that the evolutionary position of *D. discoideum* is located before the branching of metazoa and fungi but after the divergence of the plant kingdom[7], placing it close to the

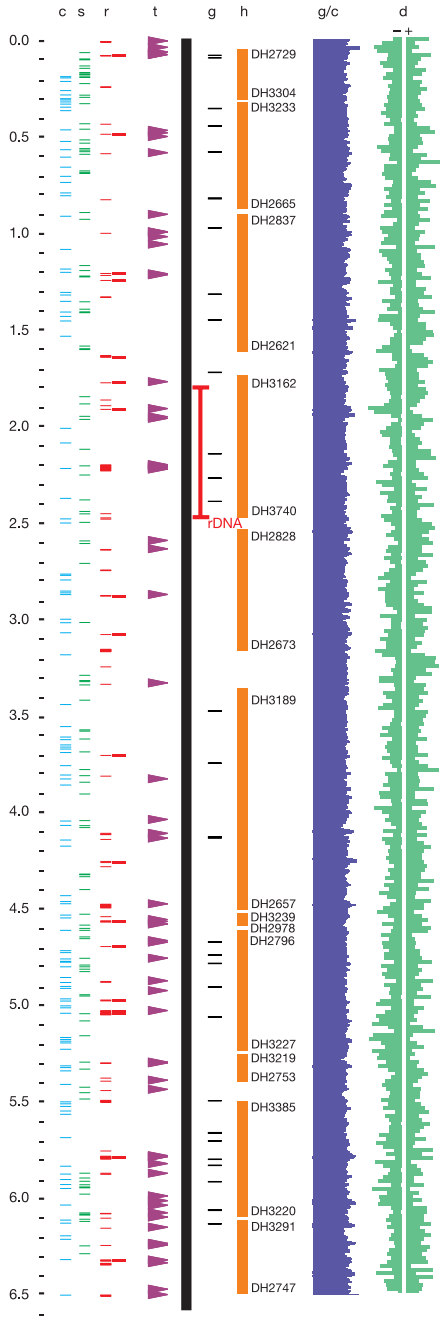


Figure 1 Feature distribution on chromosome 2. Only the linked portion (6.5 Mb) is shown (solid black line). Clone gaps (c), sequence gaps (s), repeat elements (r; heavier bars to their right indicate unresolvable clusters), tRNA genes (t) and genes (g) used to seed assembly are shown. HAPPY linkage groups (h) were used to guide assembly; only the endmost markers in each group are named. G+C content (g/c), strand-specific coding sequence density (d), the ribosomal DNA copy, and the duplicated region above it (represented here as a single copy) are shown. The centromere and telomere are respectively above and below the portion shown. An expanded version is at http://genome.imb-jena.de/dictyostelium/chr2/Chr2map.html.

base of metazoan evolution.

The natural habitat of *D. discoideum* is deciduous forest soil where the amoeboid cells feed on bacteria by phagocytosis and multiply by equal mitotic division. Exhaustion of the food source triggers a developmental programme, in which more than 100,000 cells aggregate by chemotaxis to form a multicellular structure. Morphogenesis and cell differentiation then culminate in the production of spores, enabling the organism to survive unfavourable conditions[8]. *D. discoideum* therefore lies at the borderline between free-living cells and multicellular organisms, making it ideal for the study of cellular differentiation and integration. Its haploid genome, ease of culture and genetic manipulability make it amenable to biochemical, genetic, and cell-biological approaches[9]. This allows the dissection of the molecular basis of the most fundamental cellular processes: differentiation, signal transduction, phagocytosis, cytokinesis, cell motility and chemotaxis[10–12].

To provide the basis for genome-wide investigations an international effort was initiated[13] to sequence the ~34-Mb genome of *D. discoideum*, strain AX4. Besides six chromosomes ranging from 4 to 8 Mb (refs 14, 15), the nucleus harbours approximately 100 copies of a ~90-kb palindromic chromosome containing the ribosomal RNA genes. The high A+T content (78%, exceeded only by *Plasmodium falciparum* at 80%; refs 16, 17) coupled with a high density of repetitive elements, posed severe challenges for genome sequencing. To reduce the complexity of the assembly task, the genome was analysed chromosome by chromosome, using a whole chromosome shotgun (WCS) approach. The chromosomal libraries were only ~50% pure and contained clones derived from other chromosomes, so we developed an iterative and integrated assembly strategy. This allowed us to identify contiguous DNA sequences (contigs) originating from chromosome 2 and to bridge difficult sequences. Briefly (see Methods), nonrepetitive reads from the chromosome 2-enriched libraries were binned with those from the other WCS projects, and sequences of known chromosome 2 genes were used as 'seeds' around which to build contigs. These were extended using sequence data and supplemented using read-pair information and BLAST (http://blast.wustl@adu/) analysis. To confirm the chromosomal assignment of these contigs we used the relative frequencies of the constituent sequences in the chromosomally enriched libraries of the various WCS projects.

The high A+T content, the existence of many repetitive elements and the fact that clones larger than about 5 kb were unstable in *Escherichia coli*[18,19], precluding the use of large-insert bacterial clones as second-source templates, led to three types of gaps. The first type could not be spanned by plasmid clones ('clone gaps'), presumably owing to the instability of some of the intergenic regions, which have A+T contents of up to 98%. The second type arose from clusters of repetitive elements, which could not be unambiguously resolved ('repeat gaps'). The third type ('sequence gaps') were spanned by clones which, owing to their content of long homopolymer runs (even more abundant and longer than in *P. falciparum*) or lack of targets for custom primers, were recalcitrant to repeated attempts at sequencing. Contigs divided by sequence gaps were linked by read-pair information to produce larger 'scaffolds' with a total size of 7.5 Mb. The majority of these scaffolds were then connected, oriented and their internal structure validated by using mapped genes, circular yeast artificial chromosomes (cYACs) and HAPPY map[20] data. This yielded a 'linked portion' spanning 6.5 Mb of the chromosome (Fig. 1; Table 1; http://genome.imb-jena.de/dictyostelium/chr2/Chr2map.html). Although many

Table 1 Features of chromosome 2

Feature	Value
Calculated total length (Mb)*	8.0–8.1
Total length of sequence contigs (Mb)*	7.52
Cumulated length of 71 small orphan unlinked contigs (Mb)	0.4
Number of loci containing complex repetitive elements**	58
Resolved loci	40
Unresolved loci	18
Number of tRNAs	73
Genes*†	
Predicted number	2,799
Density	1 gene/2.6 kb
Average length (bases)	1,626
Number of genes with ESTs	1,120 (40%)
AT content (%)	
Exons	72
Introns	87
Intergenic	86
Whole chromosome	77.8
Exons (coding)	
Number	6,398
Average exon number/gene	2.29
Average size (bases)	711
Introns	
Number	3,587
Average size (bases)	177
Intergenic regions	
Average size (bases)	786
Intronless genes (%)	893 (32)

*Excluding duplication of 0.7 Mb.
**In 6.5-Mb linked portion of chromosome 2.
†Excluding genes coded for in repeat loci.

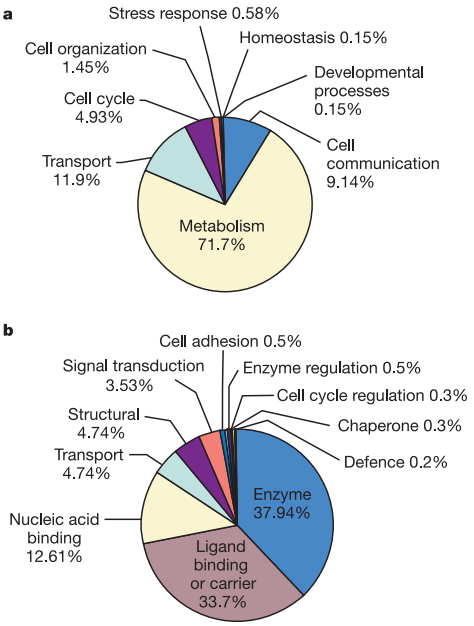


Figure 2 Functional classification of *D. discoideum* chromosome 2-coded proteins. We used the GO terminology (http://whitefly.lbl.gov/annot/go/database/index.html) for the automated classification of proteins in process (**a**) and function (**b**) groups according to their InterPro domains. The process groups contain 689 proteins, the function groups 991 proteins. Proteins with InterPro domains but no GO assignment (424) or proteins without Interpro domains (1,319) were not characterized. Currently no *D. discoideum*-specific GO terms are defined, thus leaving some of the functionally characterized *D. discoideum*-specific genes unclassified.

of the sequence and clone gaps have been closed, those that remain (95 sequence gaps and 89 clone gaps, totalling an estimated 150 kb) appear intractable. The most resistant gaps have been those containing the most A+T-rich DNA, and are hence least likely to contain sequences of biological relevance.

D. discoideum chromosomes have been reported to be acro- or telocentric with the centromere embedded in a large cluster of long terminal repeat retrotransposons (DIRS-1) composed of more than 40 elements[15,19]. The fine structure of this cluster, which lies outside the linked portion of the chromosome shown in Fig. 1, could not be resolved because of the low polymorphism rates of the complex repetitive elements[19]. It spans up to 0.5 Mb and also contains copies from other transposon families and small repetitive elements. Overall, the number of repetitive elements in *D. discoideum* genomic DNA is high compared to *S. cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*. Chromosome 2 harbours all previously described *D. discoideum* complex repetitive elements, mainly organized in clusters of intact and truncated elements[19]. There are 58 such loci (each consisting of one or more such elements) on the linked portion of chromosome 2. The fine structure of 18 of these could not be resolved and remain as 'repeat gaps' (Fig. 1; Table 1). Altogether, we estimate that complex repetitive elements represent 10.2% (approximately 0.8 Mb) of chromosome 2, corresponding well with the estimate of 9.6% for the entire genome[19]. On the basis of the combined sizes of the sequence scaffolds, the clone and sequence gaps, and the unresolved repeat regions (including the pericentromeric region), we calculate the size of the chromosome to be 8.1 Mb.

A duplication of approximately 700 kb is thought to have occurred after the separation of the laboratory strains AX2 and AX4 (ref. 15). We detected an inverse tandem repeat of similar size between the HAPPY Map markers DH3162 and DH3740, bordered at the telomeric end by an almost complete copy of the extrachromosomal rDNA palindrome (Fig. 1). This might represent a chromosomal master copy for the generation of the extrachromosomal rDNA palindrome after sexual recombination, as in *Tetrahymena thermophila*[21]. The second copy of the duplication was excluded from calculations of chromosome length and gene number.

We find that most features of the chromosome (G+C content, coding sequence density (CDS) and complex repetitive elements) are evenly distributed over the 6.5-Mb linked portion, although the distribution of the transfer RNA genes shows a slight bias towards the telomere (Fig. 1; http://genome.imb-jena.de/dictyostelium/chr2/Chr2map.html). We used gene prediction programs and database searches to determine and annotate the 2,799 putative genes of chromosome 2. A further 124 putative genes coded for by complex repetitive elements were excluded from further analyses. *D. discoideum* genes in general have few and small introns, with an average of 1.2 introns per gene. Intron length and distribution is comparable to that of *P. falciparum* and other lower eukaryotes. The mean A+T content in exons is 72%, whereas it is 87% in introns, and 86% in intergenic regions (Table 1). This extreme compositional bias may help to delineate the introns during splicing, as has been suggested in *Arabidopsis thaliana*. In support of this hypothesis, *D. discoideum* introns do contain the canonical GT–AG dinucleotides but, unusually among fully sequenced eukaryotes, all information is confined to the intron side of the splice site[22].

Turning to the gene content, expressed sequence tags (ESTs) exist for 40% (1,120) of the predicted genes (Table 1)[23]. BLAST searches against the protein sets of completely sequenced eukaryotic genomes as well as against SwissProt and TrEMBL databases showed that 45% (1,260) of the putative *D. discoideum* genes had a match ($P < 10^{-15}$), leaving the proportion of unique genes (55%) comparable to that observed for other eukaryotes. About 53% (1,480) of the putative genes contained domains defined in the InterPro database (http://www.ebi.ac.uk/interpro)[24]; again, this proportion is comparable to other eukaryotes[6]. In total, EST, protein, and/or InterPro matches provide support for 1,960 of the 2,799 predicted

Table 2 Most frequent InterPro domains

Domain	Description	DD	SC	AT	CE	DM	HS
IPR001687	ATP/GTP-binding site motif A (P-loop)*	6.07	0.57	0.61	0.32	0.46	0.33
IPR000694	Proline-rich region	3.72	NA	NA	NA	NA	NA
IPR000561	EGF-like domain*	2.18	0.02	0.16	0.68	0.62	1.28
IPR000719	Eukaryotic protein kinase	1.93	1.91	4.07	2.34	1.79	2.64
IPR002290	Serine/threonine protein kinase	1.89	1.83	3.34	1.33	1.22	1.83
IPR001245	Tyrosine protein kinase	1.71	0.05	1.84	0.84	0.65	1.22
IPR001680	G-protein beta WD-40 repeats	1.11	1.63	1.02	0.80	1.31	1.34
IPR003593	AAA ATPase superfamily*	1.11	0.95	0.90	0.40	0.56	0.46
IPR000051	SAM (and some other nucleotide) binding motif	0.89	0.33	0.40	0.25	0.28	0.20
IPR001849	Pleckstrin homology (PH) domain	0.89	0.47	0.12	0.41	0.54	1.24
IPR002048	EF-hand*	0.86	0.26	0.85	0.65	0.93	1.15
IPR001841	RING finger	0.82	0.65	1.82	0.81	0.85	1.20
IPR002085	Zinc-containing alcohol dehydrogenase superfamily	0.82	0.34	0.15	0.06	0.07	0.08
IPR000794	Beta-ketoacyl synthase*	0.79	0.03	0.02	0.02	0.03	0.01
IPR003579	RAS small GTPases, Rab subfamily	0.79	0.15	0.23	0.15	0.21	0.22
IPR001611	Leucine-rich repeat	0.75	0.13	1.93	0.33	0.83	0.74
IPR003577	RAS small GTPases, Ras subfamily	0.75	0.05	0.00	0.06	0.07	0.10
IPR003880	Phosphopantetheine attachment site	0.75	0.10	0.21	0.15	0.28	0.08
IPR000504	RNA-binding region RNP-1 (RNA recognition motif)	0.71	0.93	0.96	0.69	1.13	1.25
IPR000873	AMP-dependent synthetase and ligase	0.71	0.18	0.17	0.17	0.25	0.16
IPR003578	RAS small GTPases, Rho subfamily	0.71	0.10	0.04	0.05	0.04	0.10
IPR000345	Cytochrome *c* family haem-binding site	0.68	0.10	0.58	0.31	0.31	0.37
IPR001227	Acyl transferase domain*	0.68	0.02	0.00	0.02	0.03	0.01
IPR001806	Ras GTPase superfamily	0.68	0.36	0.38	0.33	0.51	0.60
IPR000477	RNA-directed DNA polymerase (Reverse transcriptase)	0.64	0.08	0.50	0.46	0.09	0.14
IPR001064	Zinc-finger GCS-type	0.64	0.10	0.07	0.04	0.07	0.14
IPR002110	Ankyrin-repeat*	0.64	0.29	0.44	0.53	0.62	0.91
IPR001601	Generic methyl-transferase	0.61	0.10	0.22	0.06	0.07	0.06
IPR001410	DEAD/DEAH box helicase	0.57	1.19	0.53	0.44	0.54	0.52
IPR002106	Aminoacyl-transfer RNA synthetases class-II	0.57	0.36	0.35	0.59	0.41	0.20

Occurrence of the thirty most frequent InterPro domains on *D. discoideum* chromosome 2 (DD; including repetitive elements) and fully sequenced eukaryotes. The percentage of genes in each organism that contain the respective domain type is given. SC, *S. cerevisiae*; AT, *A. thaliana*; CE, *C. elegans*; DM, *D. melanogaster*; HS, *H. sapiens*. The data for SC, CE, DM, HS and AT were taken from http://www.ebi.ac.uk/proteome/. NA, not analysed.
*These entries are discussed in the text.

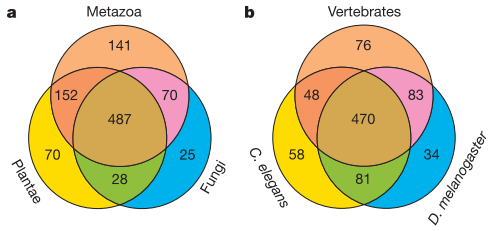


Figure 3 Phylum-specific distribution of proteins. **a**, For comparison with the chromosome 2 translated genes, plants are represented by the full protein set of *A. thaliana* and fungi by *S. cerevisiae* plus *S. pombe*. Metazoans are represented by *D. melanogaster*, *C. elegans*, the (as yet incomplete) *Homo sapiens* protein set, plus annotated nonhuman vertebrate proteins from the SwissProt database. **b**, Distribution of *D. discoideum* genes with $P < 10^{-15}$ between fully sequenced metazoan species. The vertebrate gene set is as in **a**.

genes. Applying the gene ontology (GO) terminology for the automated classification of proteins (http://whitefly.lbl.gov/annot/go/database/index.html) we could attribute functions and/or processes to 1,026 (37%) of the predicted protein products (Fig. 2). Slightly more of these proteins could be classified to functions (991) than to process (689) categories, and 654 out of the 1,026 classified proteins are present in both categories. Forty-seven per cent of the putative proteins remain unclassified and a further 15% of all proteins could not be categorized because their corresponding InterPro domains are not yet assigned to GO terms (Fig. 2).

Because *D. discoideum* undergoes differentiation and development we might expect a significant number of genes associated with multicellular life. In fact, a remarkably high proportion of GO-classified proteins are grouped into the cell communication category (9.14%) and involved in signal transduction or cell adhesion, or comprise cytoskeletal proteins containing signalling domains. As expected, analysis of InterPro matches reveals that domains required for cell motility, signalling, surface attachment and cytoskeletal functions are considerably more abundant than in yeast. When we compare the most frequent InterPro domains of chromosome 2 genes to those of other species (Table 2), the ATP/GTP-binding site motif A (P-loop) is strongly over-represented (6.07% of the predicted genes on the *D. discoideum* chromosome 2 carry this motif, versus 0.33% in human), whereas the epidermal growth factor (EGF)-like domain is over-represented only slightly compared to human (2.18% versus 1.28%), but strongly in comparison to yeast (0.02%). The AAA ATPase superfamily domain is found in comparable proportions in *D. discoideum*, *S. cerevisiae* and *A. thaliana*, but is less abundant in *C. elegans*, *D. melanogaster* and human, whereas the proportions of the Ca^{2+}-binding EF-hand domain and the ankyrin repeat are roughly comparable in all organisms with the exception of yeast, where they are less abundant. We have also identified many beta-ketoacyl synthase and acyl transferase domains, which are hardly present in the other organisms considered here. In *D. discoideum* many of these domains are part of polyketide synthases, which are exceptionally large, multifunctional proteins, primarily present in actinomycetes, bacilli and filamentous fungi. The compounds built via the polyketide synthase pathways might enable *D. discoideum* to defend itself against its natural competitors.

Many *D. discoideum* genes—particularly those involved in signalling and cell movement—are known to be present as multiple copies or as members of large gene families. This is supported by our analysis of chromosome 2, which contains 130 genes present as two or more copies ($P < 10^{-30}$ and sequence similarity over the complete length), amounting to 337 (12%) of the predicted genes. Because paralogues on the other chromosomes have not been taken into account, the number of singletons will further decrease when all chromosomes have been analysed. We have found ten genes for members of the Ras-related small GTP-binding protein family and nine genes sharing the RasGEF domain. Furthermore, we identified another G protein with homology to Gα2, a component of the cyclic AMP signalling system, and also residing on chromosome 2. We have also found two more members of the G-protein coupled receptor family. Surprisingly, these proteins have highest homology to GABA (γ-aminobutyric acid) receptors, which have not yet been found outside the metazoan branch. Genes coding for components of the cytoskeleton are frequently present in multiple copies. Of the ~27 actin genes (including pseudogenes) in the *D. discoideum* genome[19], thirteen are present on chromosome 2 and ten of these translate into identical protein sequences. Chromosome 2 harbours several genes coding for motor proteins, among which are six genes for different unconventional myosins. Although the cytoskeleton has been intensively studied, we have found putative new paralogues for profilin I/II, fimbrin, cofilin 1/2 and the unconventional myosin gene family. The discovery of additional putative paralogues of cytoskeletal proteins supports the concept of functional redundancy in the cytoskeletal system[12]. The ABC (ATP-binding cassette) transporter family is probably one of the largest in the genome. There are thirteen such genes on chromosome 2, including several members of the ABC A subfamily whose occurrence has been restricted to multicellular eukaryotes. ABC transporters use the energy of ATP hydrolysis to translocate specific substrates across cellular membranes. Mutations in many of the human genes coding for ABC transporters are associated with disease such as cystic fibrosis, Stargardt's disease or hyperinsulinism. We have found genes on chromosome 2 with high similarities to the

Table 3 *D. discoideum* chromosome 2 genes with similarity to human disease genes

Disease (gene symbol)	OMIM number	Accession number	*D. discoideum* gene	BLASTP value ($<1.0 \times 10^{-50}$)
Renal tubular acidosis (ATP6B1)	192132	AAD11943	dd_01070	4.0e-246 (0)
Immunodeficiency (DNA Ligase 1)	126391	NP_000225	dd_02463	1.9e-245 (0)
Hereditary nonpolyposis colorectal cancer, type 1 (HNPCC) (MSH2)	120435	AAA18643	dd_00995	2.4e-237 (1)
*Hyperinsulinism (ABCC8)	600509	Q09428	dd_00006	3.0e-220 (1)
G6PD deficiency (G6PD)	305900	NP_000393	dd_01534	5.1e-190 (0)
*Stargardt's (ABCA4)	601691	AAC51144	dd_02412	6.5e-189 (3)
Deafness, hereditary (MYO15)	602666	AAF05903	dd_02568	1.2e-182 (5)
Familial cardiac myopathy (MYH7)	160760	P12883	dd_02401	4.2e-177 (5)
Chediak–Higashi (CHS1)	214500	NP_000072	dd_02608	3.3e-151 (1)
Cancer (AKT2)	164731	AAA58364	dd_02928	1.5e-94 (2)
HNPCC (MSH3)	600887	AAB06045	dd_01030	8.9e-78 (1)

From a list of 287 confirmed human disease protein sequences[30] those are shown that match a *D. discoideum* chromosome 2 protein with a BLASTP probability of less than 1.0×10^{-50}, indicating a strong similarity. Only the best hit is listed and the total number of additional strong hits ($P < 1.0 \times 10^{-50}$) is given in parentheses after the probability score. OMIM, Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim/).
*Homologous proteins discussed in the text.

latter two genes and to several other human disease-related genes (Table 3).

What can the genome of *D. discoideum* tell us about the common genomic repertoire of eukaryotic life? To address this question, we compared the protein products of the 2,799 genes of chromosome 2 to the complete protein sets of fully sequenced eukaryotes ($P < 10^{-15}$) and found that 973 proteins (35%) have matches. Of these only 487 share similarities across plants (*A. thaliana*), fungi (*S. cerevisiae* and *S. pombe*) and metazoa (*C. elegans*, *D. melanogaster* and available vertebrate sequences). A surprisingly high number (141) have matches with metazoa but not plants or fungi (Fig. 3a). Subdividing metazoa into fly, worm and vertebrates shows that even amongst this closely related group not all genes have comparable similarities in each species (Fig. 3b). This may reflect gene losses during evolution, or evolutionary rate variations for identical genes in different organisms, but could also reflect a gain of function for specific gene groups in each organism. If chromosome 2 is taken as a representative quarter of the *D. discoideum* genome, then less than 2,400 different genes are shared between *D. discoideum*, *S. pombe*, *S. cerevisiae*, *C. elegans*, *A. thaliana*, *D. melanogaster* and man. This number might well represent the 'minimal gene set' of a free-living eukaryote. From random mutagenesis studies it was previously estimated that the essential genes of yeast comprise only about 30% of all its genes[25]. Our estimate of the number of genes shared by all eukaryotes is close to this number.

Our analyses are all predicated on the assumption that chromosome 2, representing 25% of the genome, is typical of the remainder. This seems a reasonable assumption and other evidence on the distribution of mapped genes[15] does not suggest that chromosome 2 is particularly atypical. Our findings can also clarify the evolutionary position of *D. discoideum*. Its genome exhibits greater similarities to metazoa than to plants or fungi (Fig. 3). This supports the finding of a recent phylogenetic analysis of conserved protein sequences which placed the Myxomycota (to which *D. discoideum* belongs) at a position before the branching of the metazoa and fungi but after the divergence of the plant kingdom[7]. *D. discoideum* does not appear to have suffered the extensive gene loss observed in *S. cerevisiae* and therefore its gene content may better represent a basic eukaryotic genome. This conservation of the complete gene set makes *D. discoideum* well suited for functional studies of genes not represented in yeast. Its surprisingly high gene number may in part reflect the higher order of complexity associated with multicellular life. □

Methods

Further information on sequence data and analysis results can be accessed via http://genome.imb-jena.de/dictyostelium/ and http://www.uni-koeln.de/dictyostelium/.

Sequencing and assembly

Library construction and sequencing was done as described previously[19]. 160,000 chromosome 2 library-derived reads were pooled with the nonrepetitive reads from other *D. discoideum* whole chromosome shotgun projects to give a total of 500,000 reads. The subset of reads matching genes mapped to chromosome 2 (ref. 15) were assembled to build seed contigs, and further contigs were assembled around complementary reads from the clones in these seeds (for details see http://genome.imb-jena.de/dictyostelium/chr2/seeds.html). The previously published[15] map order of the 'seed' genes was largely confirmed. Considering the combined data of the HAPPY map and sequence assembly, it is likely that the discrepancies arise from errors in the earlier YAC contigs, which have been shown to suffer from a proportion of misplaced clones[26]. The assembly database was then enlarged by the incorporation of reads with a higher than average frequency of occurrence in the chromosome 2 library reads (these are more likely to originate from chromosome 2 than from other chromosomes, owing to our preferential use of the chromosome 2 specific clone libraries). The contigs were extended by the incorporation of further reads which were found by BLAST analysis of the contig ends. This assembly method yielded about 1,100 contigs larger than 2 kb. Chromosomal assignment of each contig was checked on the basis of its content of sequences derived from each of the different chromosome-enriched libraries (K.S., unpublished software). In this way, contaminant sequences were filtered out; conversely, reads derived from the other whole chromosome shotgun projects but assigned to chromosome 2 were incorporated into our assembly. To ensure that we had not missed portions of the chromosome by this strategy we assembled all chromosome 2 library-derived reads and checked the resulting contigs for chromosome specificity. The resulting additional contigs were added to the chromosome 2-specific assembly. All contigs were manually inspected to ensure data accuracy. Clones spanning sequencing gaps between neighbouring contigs defined scaffolds. Directed closure of these gaps was done using custom primers to walk on existing clones. Additional gaps were closed by using the transposon insertion technique and polymerase chain reaction (PCR) approaches.

Mapping

As part of the ongoing genome-wide *D. discoideum* HAPPY mapping project, a short range (~100 kb), high-resolution (mean, 15 kb) HAPPY mapping panel was prepared from AX4 genomic DNA, pre-amplified by PEP (primer extension pre-amplification) and diluted before use as a template for marker typing. Hemi-nested primers were designed for 824 markers selected from the sequencing projects. Markers were typed onto the HAPPY mapping panel and sorted into linkage groups as previously described[26]. Maps for each linkage group were generated and validated by inspection to reduce the risk of incorporating spurious intermarker linkages. The results obtained so far define the order of 365 chromosome 2 markers in 12 large linkage groups (for details see http://genome.imb-jena.de/dictyostelium/chr2/linkage.html). Further positional information was obtained from PCR screening of a cYAC library with insert sizes of 80–100 kb covering the genome approximately sevenfold. Sequence contigs or HAPPY linkage groups were assumed to be linked if primer pairs derived from them hit the same cYAC clone(s). By integrating the HAPPY, cYAC and sequence scaffold data, a region spanning 6.5 Mb was robustly assembled. Only four sequence scaffolds, totalling 0.6 Mb, could not be placed onto the map. Of these, two are too large to fit in the gaps of the linked 6.5-Mb portion of the chromosome, and are presumably located at the ends. The assembly produced 71 unlinked orphan contigs amounting to 0.41 Mb, consisting mainly of fragments of complex repetitive elements.

Sequence analysis

G+C content was calculated using a sliding window of 10,000 bases and a step size of 1,000 bases. Strand-specific CDS density was measured as percentage of coding triplets in a stepped window of 5,000 bases. A database containing the complex repetitive elements of *D. discoideum* was used with RepeatMasker to scan the sequence for repeats[19]. tRNAs were detected using tRNAscan-SE[27]. To define the genes on chromosome 2, the gene prediction program GeneID[28] was trained with 140 known *D. discoideum* genes and its parameters adjusted to be able to define proper gene borders and intron positions. The lower limit for the gene length was 120 bases of coding sequence. EST matches were defined by BLAST with >98% identity, and word length of 32. The protein products of predicted genes were compared to the databases of completed genomes: The *Arabidopsis* Information Resource (http://www.arabidopsis.org/home.html), Wormpep (http://www.sanger.ac.uk/Projects/C_elegans/wormpep/), ftp://ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/, *Saccharomyces* Genome Database (http://genome-www.stanford.edu/Saccharomyces/), The *Schizosaccharomyces pombe* Genome Sequencing Project (http://www.sanger.ac.uk/Projects/S_pombe/), Ensembl Genome Browser (http://www.ensembl.org) as well as against SWISS-PROT entries and TrEMBL. They were also checked for the presence of InterPro domains using the InterPro database (http://www.ebi.ac.uk/interpro). Functional classification was done automatically using the GO classification system (http://www.geneontology.org/)[29].

Received 14 December 2001; accepted 26 April 2002; doi:10.1038/nature00847.

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546–567 (1996).
2. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).
3. The *C. elegans* Sequencing Consortium Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
4. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
5. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
6. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
7. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290,** 972–977 (2000).
8. Loomis, W. F. Genetic networks that regulate development in *Dictyostelium* cells. *Microbiol. Rev.* **60,** 135–150 (1996).
9. Eichinger, L., Lee, S. S. & Schleicher, M. *Dictyostelium* as model system for studies of the actin cytoskeleton by molecular genetics. *Microsc. Res. Technol.* **47,** 124–134 (1999).
10. Parent, C. A. & Devreotes, P. N. A cell's sense of direction. *Science* **284,** 765–770 (1999).
11. Firtel, R. A. & Meili, R. *Dictyostelium*: a model for regulated cell movement during morphogenesis. *Curr. Opin. Genet. Dev.* **10,** 421–427 (2000).
12. Noegel, A. A. & Schleicher, M. The actin cytoskeleton of *Dictyostelium*: a story told by mutants. *J. Cell Sci.* **113,** 759–766 (2000).
13. Kay, R. R. & Williams, J. G. The *Dictyostelium* genome project: an invitation to species hopping. *Trends Genet.* **15,** 294–297 (1999).
14. Cox, E. C., Vocke, C. D., Walter, S., Gregg, K. Y. & Bain, E. S. Electrophoretic karyotype for *Dictyostelium discoideum*. *Proc. Natl Acad. Sci. USA* **87,** 8247–8251 (1990).
15. Loomis, W. F. & Kuspa, A. Dictyostelium—*A Model System for Cell and Developmental Biology* (eds Maeda, Y., Inouye, K. & Takeuchi, I.) 15–30 (Universal Academic, Tokyo, 1997).
16. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
17. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
18. Glöckner, G. Large scale sequencing and analysis of AT rich eukaryote genomes. *Curr. Genom.* **1,** 289–299 (2000).

19. Glöckner, G. *et al.* The complex repeats of *Dictyostelium discoideum*. *Genome Res.* **11,** 585–594 (2001).
20. Dear, P. H. in *Genome Mapping—A Practical Approach* (ed. Dear, P. H.) 95–124 (IRL Press, Oxford, 1997).
21. Pan, W. C. & Blackburn, E. H. Single extrachromosomal ribosomal RNA gene copies are synthesized during amplification of the rDNA in *Tetrahymena*. *Cell* **23,** 459–466 (1981).
22. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA* **98,** 11193–11198 (2001).
23. Morio, T. *et al.* The *Dictyostelium* developmental cDNA project: generation and analysis of expressed sequence tags from the first-finger stage of development. *DNA Res.* **5,** 335–340 (1998).
24. Apweiler, R. *et al.* InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16,** 1145–1150 (2000).
25. Goebl, M. G. & Petes, T. D. Most of the yeast genomic sequences are not essential for cell growth and division. *Cell* **46,** 983–992 (1986).
26. Konfortov, B. A., Cohen, H. M., Bankier, A. T. & Dear, P. H. A high-resolution HAPPY map of *Dictyostelium discoideum* chromosome 6. *Genome Res.* **10,** 1737–1742 (2000).
27. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
28. Parra, G., Blanco, E. & Guigo, R. GeneID in *Drosophila*. *Genome Res.* **10,** 511–515 (2000).
29. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
30. Fortini, M. E., Skupski, M. P., Boguski, M. S. & Hariharan, I. K. A survey of human disease gene counterparts in the *Drosophila* genome. *J. Cell Biol.* **150,** F23–F30 (2000).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

Acknowledgements

We thank S. Förste, N. Zeisse, S. Rothe, S. Landmann, R. Schultz, S. Müller and R. Müller for expert technical assistance. We also thank the working team of the Japanese cDNA project (http://www.csm.biol.tsukuba.ac.jp/cDNAproject.html) for sharing data. The sequencing of chromosome 2 was supported by the Deutsche Forschungsgemeinschaft, with partial support by Köln Fortune. Additional support was obtained from the NIH, the Medical Research Council and the EU.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.A.N. (e-mail: noegel@uni-koeln.de) or G.G. (e-mail: gernot@imb-jena.de) or L.E. (e-mail: ludwig.eichinger@uni-koeln.de).

|| The *Dictyostelium* Genome Sequencing Consortium (members not included in the main author list):

Sequencing and Analysis:
The Sanger Institute Dictyostelium sequencing team (led by Bart G. Barrell & Marie-Adèle Rajandream)[1], Jeffrey G. Williams[2], Robert R. Kay[3], Adam Kuspa[4], Richard Gibbs[4], Richard Sucgang[4], Donna Muzny[4] & Brian Desany[4]

Generation of cYAC library:
Kathy Zeng[5], Baoli Zhu[5] & Pieter de Jong[5]

Advisory Committee for the DFG-funded project:
Theodor Dingermann[6], Günther Gerisch[7], Peter Philippsen[8], Michael Schleicher[9], Stephan C. Schuster[10] & Thomas Winckler[6]

1, The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; 2, University of Dundee, MSI/WTB Complex, Dundee, UK; 3, MRC Laboratory, of Molecular Biology, Cambridge CB2 2QH, UK; 4, Baylor College of Medicine, Houston, Texas 77030, USA; 5, Children's Hospital Oakland – BACPAC Resources, Oakland, California 94609, USA; 6, Institut für Pharmazeutische Biologie, Universität Frankfurt (Biozentrum), Frankfurt am Main, 60439, Germany; 7, Max-Planck-Institut für Biochemie, 82152 Martinsried, Germany; 8, Molecular Microbiology, Biozentrum der Universität, 4056 Basel, Switzerland; 9, A.-Butenandt-Institut/Zellbiologie, Ludwig-Maximilians-Universität, 80336 München, Germany; 10, Max-Planck-Institut für Entwicklungsbiologie, 72076 Tübingen, Germany

Intracellular calcium stores regulate activity-dependent neuropeptide release from dendrites

Mike Ludwig*, Nancy Sabatier*, Philip M. Bull*, Rainer Landgraf†, Govindan Dayanithi‡ & Gareth Leng*

** Department of Biomedical Sciences, University of Edinburgh Medical School, George Square, Edinburgh EH8 9XD, UK*
† Max Planck Institute of Psychiatry, Clinical Institute, Kraepelinstraße 2-10, 80804 Munich, Germany
‡ Department of Neurobiology, INSERM 432, University of Montpellier II, Place Eugene Bataillon, F-34094 Montpellier, Cedex 5, France

Information in neurons flows from synapses, through the dendrites and cell body (soma), and, finally, along the axon as spikes of electrical activity that will ultimately release neurotransmitters from the nerve terminals. However, the dendrites of many neurons also have a secretory role, transmitting information back to afferent nerve terminals[1–4]. In some central nervous system neurons, spikes that originate at the soma can travel along dendrites as well as axons, and may thus elicit secretion from both compartments[1]. Here, we show that in hypothalamic oxytocin neurons, agents that mobilize intracellular Ca^{2+} induce oxytocin release from dendrites without increasing the electrical activity of the cell body, and without inducing secretion from the nerve terminals. Conversely, electrical activity in the cell bodies can cause the secretion of oxytocin from nerve terminals with little or no release from the dendrites. Finally, mobilization of intracellular Ca^{2+} can also prime the releasable pool of oxytocin in the dendrites. This priming action makes dendritic oxytocin available for release in response to subsequent spike activity. Priming persists for a prolonged period, changing the nature of interactions between oxytocin neurons and their neighbours.

Neurons in the supraoptic nucleus (SON) of the hypothalamus project axons to the posterior pituitary, where oxytocin and vasopressin are secreted from axonal nerve terminals into the systemic circulation. These peptides are also released in large amounts from dendrites in the SON[5], but secretion at these two sites is not consistently correlated. Suckling evokes oxytocin release in the SON[6] before significant peripheral secretion, whereas after osmotic stimulation, SON oxytocin release lags behind peripheral secretion[7]. During lactation, in response to suckling, oxytocin cells discharge with brief, intense bursts[8]; these bursts release boluses of oxytocin into the circulation that result in milk let-down from the mammary glands. The bursting activity can be blocked by central administration of oxytocin antagonists[9], thus central as well as peripheral oxytocin is essential for milk let-down. It has been proposed that suckling evokes dendritic oxytocin release that acts in a positive feedback manner to evoke bursting[10].

Oxytocin mobilizes intracellular Ca^{2+} from thapsigargin-sensitive stores in oxytocin cells[11]. Here we tested the hypothesis that this might be critical for dendritic oxytocin release. In anaesthetized rats, we implanted a microdialysis probe into the SON to measure oxytocin release in response to systemic osmotic stimulation. In some of these experiments we applied thapsigargin directly to the SON through the dialysis probe. Thapsigargin caused a significant increase in SON oxytocin release that returned to control levels after washout. Subsequent systemic osmotic stimulation (2 ml of 1.5 M NaCl, intraperitoneal injection) caused a much larger release of oxytocin in thapsigargin-pretreated rats than in controls. Osmotically stimulated oxytocin secretion into the circulation was unaffected by exposure of one SON to thapsigargin (Fig. 1a, b).

To test whether thapsigargin potentiated spike-dependent release

Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotipe

Tetraodon Genome Sequencing Consortium (including G. Parra and R. Guigó).
Nature (431):946-957 (2004)

The fish *Tetraodon nigroviridis* lives in the rivers and estuaries of Indonesia, Malaysia and India. As a vertebrate, its gene pool is very similar to that of other vertebrates, including mammals such as humans and mice. It has been observed that the genome of *T. rubripes*, another puffer fish from the same family, has a remarkably low content of repetitive DNA, and this also applies to *T. nigroviridis*. Therefore, for geneticists interested in studying genes, fishes of the Tetraodontiform family have a huge advantage over mammals: their gene pool is contained within approximately eight times less DNA (i.e. the genome is eight times smaller). This feature allows us to rapidly target our studies on the interesting part: the genes.

Because of the compactness of the genome, the fraction of intergenic and intronic regions is smaller than in the other vertebrates. Thus, the identification of the structure of coding genes should be easier than in other vertebrates. This should allow, for the first time, to get the complete picture of the gene content of a vertebrate genome. This collection should also serve as a reference for comparisons with the human genome.

Our participation was focus on the gene annotation process (section 5 of the presented article). A `geneid` parameter file was obtained based on a set of annotated sequences. GENOSCOPE, the center that was in charge of the *Tetraodon* genomic annotation, provided us with a set of 10 sequences containing 117 *Tetraodon* genes. From these sequences we built a parameter file for `geneid` that was used for the final annotation of the *T. nigroviridis* genome.

Building the parameter file

The data set annotation was checked before the extraction of the biological information. From the 117 annotated genes. One had a mis-annotated exon indicating that the end of the annotation was not clear. Four CDSs contained stop codons in frame and four did not contain the ATG start site. Finally, we gathered split the contig sequences in 108 sequences containing the validated genes for the training set. The sequences contained 290,567 coding nucleotides and 150,345 intronic nucleotides.

The data set contained 922 exons, 922 acceptor and donor sites and 108 translational start sites. The amount of information of every position was analyzed for each signal. The splice sites seemed to not differ substantially from the other vertebrate splice sites (see Figure 3.3).

PWMs were computed to model the splice sites and the translational start signal. For the start site modeling, the six positions upstream of the beginning and six positions downstream were selected. For the donor and acceptor sites position from -3 to 7 and from -18 to 2 were considered informative (with 1 being position after the cleavage of the splice site). The Markov Models of order 5 and order 4 were computed using the cod-

ing regions of the 108 sequences and considering the intronic regions as the background model.

At this point, an optimization of *EW* with the new parameter file was done. The optimal *EW* was found to be -9. The results of `geneid` with *Tetraodon* parameters file showed a higher sensitivity than either `geneid` run using human parameters (see Table 7.2).

Benchmarking Tetraodon gene predictions

The same training set was first used to test the accuracy of the new parameter files. And apparently the parameter file with the Markov Model of order 5 had higher accuracy than the one using a Markov Model of order 4 (see Table 7.2). Running `geneid` with *Tetraodon* parameter file had the best performance although specificity was low (70% at nucleotide level).

To analyze a possible over-training of the parameter files, we developed a novel testing procedure based on the classical 'leave-one-out' or Jack-knife protocol. This procedure consists in leaving one example out at a time from the training set; repeating the training with the rest of sequences to build a new parameter file; testing the accuracy on the single example. Thus, the final accuracy is the average of the individual accuracy values computed in every leave-one-out round.

The results after Jack-knife protocol shows that accuracy decrease in the predictions obtained in both parameter file (the one trained with a Markov Model of order 5 and 4). However, the lost of accuracy is more pronounced in the parameter file generated with a Markov Model of order 5. This phenomenon can be explained because higher order Markov Models need larger amount of parameter to be estimated. Thus the lost of accuracy is due the over estimation of the statistical model. In order to create a matrix for order n, at least $90 * 4^{n+1}$ bases of CDS and $30 * 4^{n+1}$ bases of non-coding sample sequence are required, as estimated by Mark Borodovsky (personal communication). Therefore, for a Markov Model of order 5 at least 368,640 coding bases are needed. Using the Jack-knife approach we have confirmed that smaller samples will generate less accurate predictions. The parameter file generated with the Markov Model of order 4 was finally selected for the annotation of the *T. nigroviridis* genome.

Final annotation of the Tetraodon genome

The way to obtain the final annotation gene sets was using a combination of different sources of information: `geneid` and `genscan`, as *ab initio* gene finding programs, exofish regions, `genewise` and `est_genome` alignments. All this genomic features were integrated using `gaze`, a generic framework for the integration of gene-prediction data by dynamic programming.

The following pages correspond to the article describing the properties of the *Tetraodon nigroviridis* genome. It is also included the section 5 of the supplementary materials that corresponds to the test and the results of the combination of `geneid` and `genscan` in to the `gaze` pipeline.

	Base level			Exon level				
	Sn	Sp	CC	Sn	Sp	SnSp	ME	WE
Tetraodon MM5 -self-	0.94	0.87	0.88	0.71	0.68	0.70	0.08	0.12
Tetraodon MM5 -jkf-	0.74	0.81	0.72	0.50	0.57	0.54	0.26	0.18
Tetraodon MM4 -self-	0.92	0.80	0.82	0.66	0.61	0.64	0.10	0.18
Tetraodon MM4 -jkf-	0.87	0.79	0.78	0.60	0.58	0.59	0.15	0.19
human	0.80	0.76	0.72	0.49	0.59	0.54	0.26	0.15

Table 7.2: Accuracy of `geneid` using different parameter files and with the human profile. Tested on 108 *Tetraodon* sequences flanked by 1000 bp at both sides of the coding region. Markov coding matrices of order 5 (MM5) and 4 (MM4) were used in developing the parameter file. The evaluations have been done using the same training and test set for the -self- group and using the Jack-knife protocol in the -jkf- group.

articles

Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype

Olivier Jaillon[1], Jean-Marc Aury[1], Frédéric Brunet[2], Jean-Louis Petit[1], Nicole Stange-Thomann[3], Evan Mauceli[3], Laurence Bouneau[1], Cécile Fischer[1], Catherine Ozouf-Costaz[4], Alain Bernot[1], Sophie Nicaud[1], David Jaffe[3], Sheila Fisher[3], Georges Lutfalla[5], Carole Dossat[1], Béatrice Segurens[1], Corinne Dasilva[1], Marcel Salanoubat[1], Michael Levy[1], Nathalie Boudet[1], Sergi Castellano[6], Véronique Anthouard[1], Claire Jubin[1], Vanina Castelli[1], Michael Katinka[1], Benoît Vacherie[1], Christian Biémont[7], Zineb Skalli[1], Laurence Cattolico[1], Julie Poulain[1], Véronique de Berardinis[1], Corinne Cruaud[1], Simone Duprat[1], Philippe Brottier[1], Jean-Pierre Coutanceau[4], Jérôme Gouzy[8], Genis Parra[6], Guillaume Lardier[1], Charles Chapple[6], Kevin J. McKernan[9], Paul McEwan[9], Stephanie Bosak[9], Manolis Kellis[3], Jean-Nicolas Volff[10], Roderic Guigó[6], Michael C. Zody[3], Jill Mesirov[3], Kerstin Lindblad-Toh[3], Bruce Birren[3], Chad Nusbaum[3], Daniel Kahn[8], Marc Robinson-Rechavi[2], Vincent Laudet[2], Vincent Schachter[1], Francis Quétier[1], William Saurin[1], Claude Scarpelli[1], Patrick Wincker[1], Eric S. Lander[3,11], Jean Weissenbach[1] & Hugues Roest Crollius[1]*

[1]*UMR 8030 Genoscope, CNRS and Université d'Evry, 2 rue Gaston Crémieux, 91057 Evry Cedex, France*
[2]*Laboratoire de Biologie Moléculaire de la Cellule, CNRS UMR 5161, INRA UMR 1237, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France*
[3]*Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA*
[4]*Muséum National d'Histoire Naturelle, Département Systématique et Evolution, Service de Systématique Moléculaire, CNRS IFR 101, 43 rue Cuvier, 75231 Paris, France*
[5]*Défenses Antivirales et Antitumorales, CNRS UMR 5124, 1919 route de Mende, 34293 Montpellier Cedex 5, France*
[6]*Grup de Recerca en Informàtica Biomèdica, IMIM-UPF and Programa de Bioinformàtica i Genòmica (CRG), Barcelona, Catalonia, Spain*
[7]*CNRS UMR 5558 Biométrie et Biologie Evolutive, Université Lyon 1, 69622 Villeurbanne, France*
[8]*INRA-CNRS Laboratoire des Interactions Plantes Micro-organismes, 31326 Castanet Tolosan Cedex, France*
[9]*Agencourt Bioscience Corporation, Massachusetts 01915, USA*
[10]*Biofuture Research Group, Evolutionary Fish Genomics, Physiologische Chemie I, Biozentrum, University of Wuerzburg, Am Hubland, D-97074 Wuerzburg, Germany*
[11]*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA*

* Present address: CNRS UMR8541, Ecole Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France

Tetraodon nigroviridis **is a freshwater puffer fish with the smallest known vertebrate genome. Here, we report a draft genome sequence with long-range linkage and substantial anchoring to the 21** ***Tetraodon*** **chromosomes. Genome analysis provides a greatly improved fish gene catalogue, including identifying key genes previously thought to be absent in fish. Comparison with other vertebrates and a urochordate indicates that fish proteins have diverged markedly faster than their mammalian homologues. Comparison with the human genome suggests ~900 previously unannotated human genes. Analysis of the** ***Tetraodon*** **and human genomes shows that whole-genome duplication occurred in the teleost fish lineage, subsequent to its divergence from mammals. The analysis also makes it possible to infer the basic structure of the ancestral bony vertebrate genome, which was composed of 12 chromosomes, and to reconstruct much of the evolutionary history of ancient and recent chromosome rearrangements leading to the modern human karyotype.**

Access to entire genome sequences is revolutionizing our understanding of how genetic information is stored and organized in DNA, and how it has evolved over time. The sequence of a genome provides exquisite detail of the gene catalogue within a species, and the recent analysis of near-complete genome sequences of three mammals (human[1], mouse[2] and rat[3]) shows the acceleration in the search for causal links between genotype and phenotype, which can then be related to physiological, ecological and evolutionary observations. The partial sequence of the compact puffer fish *Takifugu rubripes* genome was obtained recently and this survey provided a preliminary catalogue of fish genes[4]. However, the *Takifugu* assembly is highly fragmented and as a result important questions could not be addressed.

Here, we describe and analyse the genome sequence of the freshwater puffer fish *Tetraodon nigroviridis* with long-range linkage and extensive anchoring to chromosomes. *Tetraodon* resembles *Takifugu* in that it possesses one of the smallest known vertebrate genomes, but as a popular aquarium fish it is readily available and is easily maintained in tap water (see Supplementary Notes for naming conventions, natural habitat and phylogeny). The two puffer fish diverged from a common ancestor between 18–30 million years (Myr) ago and from the common ancestor with mammals about 450 Myr ago[5]. This long evolutionary distance provides a good contrast to distinguish conserved features from neutrally evolving DNA by sequence comparison. *Tetraodon* sequences in fact had an important role in providing a reliable estimate of the number of genes in the human genome[6].

There has been a vigorous and unresolved debate as to whether a whole-genome duplication (WGD) occurred in the ray-finned fish (actinopterygians) lineage after its separation from tetrapods[7–9]. By exploiting the extensive anchoring of the *Tetraodon* sequence to chromosomes, we provide a definitive answer to this question. The distribution of duplicated genes in the genome reveals a striking pattern of chromosome pairing, and the correspondence of orthologues with the human genome show precisely the signatures expected from an ancient WGD followed by a massive loss of duplicated genes.

Moreover, we find that relatively few interchromosomal

Table 1 **Assembly statistics**

Parameter	Number	N50 length (kb)	Size with gaps included (Mb)	Size with gaps excluded (Mb)	Longest (kb)	Percentage of the genome with gaps included
All contigs	49,609	16	312.4	312.4	258	91.9
All scaffolds	25,773	984	342.4	312.4	7,612	100.7
All ultracontigs	128	7,622	276.4	247.0	12,035	81.3
Mapped contigs	16,083	26	197.7	197.7	258	58.1
Mapped scaffolds	1,588	608	218.4	197.7	7,612	64.2
Mapped ultracontigs	39	8,701	219.7	197.7	12,035	64.6

rearrangements occurred in the *Tetraodon* lineage over several hundred million years after the WGD. This allows us to propose a karyotype of the ancestral bony vertebrate (Osteichthyes) composed of 12 chromosomes, and to uncover many unknown evolutionary breakpoints that occurred in the human genome in the past 450 Myr.

The *Tetraodon* genome sequence

Sequencing and assembly

The *Tetraodon* genome was sequenced using the whole-genome shotgun (WGS) approach. Random paired-end sequences providing 8.3-fold redundant coverage were produced at Genoscope (GSC) and the Broad Institute of MIT and Harvard (see Supplementary Table SI1). From this, the assembly program Arachne[10,11] constructed 49,609 contigs for a total of 312 megabases (Mb; Table 1), which it then connected into 25,773 scaffolds (or supercontigs) covering 342 Mb (including gaps; see Supplementary Information). Half of the assembly is in 102 scaffolds larger than 731 kilobases (kb; the N50 length) and the largest scaffold measures 7.6 Mb, the typical length of a *Tetraodon* chromosome arm.

We produced additional data to physically link scaffolds and anchor them to chromosomes. These data include probe hybridizations to arrayed bacterial artificial chromosome (BAC) libraries, restriction digest fingerprints of BAC clones, additional linking clone sequence, alignment to available *Takifugu* sequence and two-colour fluorescence *in situ* hybridization (FISH) (see Supplementary Information). The impact of these additional mapping data was twofold: first, we could join 2,563 scaffolds in 128 'ultracontigs' that cover 81.3% of the assembly, and second, we were able to anchor the 39 ultracontigs among the largest (covering 64.6% of the assembly, with an N50 size of 8.7 Mb) to *Tetraodon* chromosomes (Fig. 1; see also Supplementary Table SI2 and Supplementary Notes).

The accuracy of the assembly was experimentally tested and the inter-contig links found to be correct in >99% of cases. On the basis of a re-sequencing experiment, we estimate that the assembly covers >90% of the euchromatin of the *Tetraodon* genome (Supplementary Information). Finally, the overall genome size was directly measured by flow cytometry experiments on several fish; an average value of 340 Mb was obtained, consistent with the sequence assembly and smaller than the previously reported estimate of 350–400 Mb.

The *Tetraodon* draft sequence has roughly 60-fold greater con-

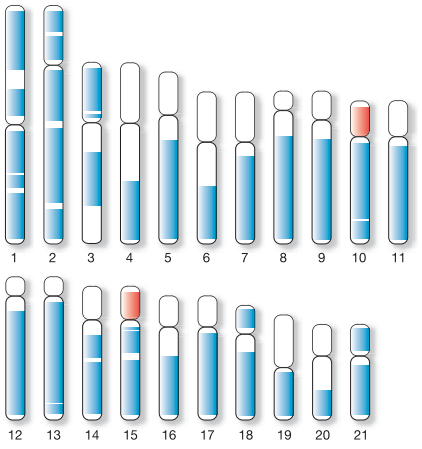


Figure 1 The *Tetraodon* genome is composed of 21 chromosomes. Red areas indicate the location of 5S and 28S ribosomal RNA gene arrays on chromosome 10 and chromosome 15, respectively. Many chromosomes are subtelocentric; that is, they only possess a very short heterochromatic arm. The extent of 39 sequence-based ultracontigs that cover about 64% of their length is shown in blue. In addition, approximately 16% of the genome is contained in another 89 ultracontigs that are not yet anchored on chromosomes, and the remaining 20% of the genome is in 23,210 smaller scaffolds.

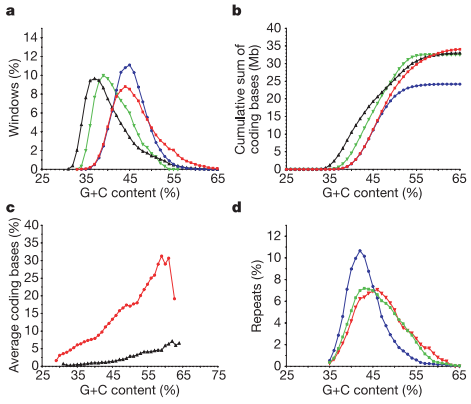


Figure 2 Distribution of the G + C content. **a**, Distribution in 5-kb non-overlapping windows across *Tetraodon* (red squares) and *Takifugu* (blue circles) scaffolds, and in 50-kb windows in human (black triangles) and mouse (green inverted triangles) chromosomes. Windows containing more than 25% ambiguous or unknown nucleotides (gaps) were excluded from the analysis. **b**, Cumulative sum of annotated coding bases in *Tetraodon* and *Takifugu* (5-kb non-overlapping windows) and human and mouse (50-kb windows) as a function of G + C content. **c**, In sharp contrast to *Takifugu*[4] the density of genes increases with the G + C content (%) in *Tetraodon* (red circles) much more than in human (black triangles). **d**, The three major families of repeats in *Tetraodon* are not distributed uniformly in the genome: long terminal repeat (LTR) and LINE elements (red diamonds and green squares, respectively) concentrate in (G + C)-rich regions and SINE elements (blue circles) concentrate in (A + T)-rich regions. In contrast, the distribution of these elements is much more uniform in *Takifugu* (Supplementary Fig. S4).

Table 2 **Comparison between *Tetraodon* and *Takifugu* annotations**

Parameter	*Tetraodon*	*Takifugu**	*Takifugu*†
Annotated genes	27,918	35,180	20,796
Annotated transcripts	27,918	38,510	33,003
Average number of coding exons per gene	6.9	4.3	8.6
Average number of UTR exons per gene	0.4	0‡	0.07
Average gene size (bp)	4,778	2,754	6,547
Average CDS size (bp)	1,230	745	1,397
Average exon size (bp)	178	171	163
Number of annotated bases (Mb)			
Coding	33.9	26.1	29.1
UTR	2.4	0‡	0.02

**Takifugu* annotations are from Ensembl version 18.2.1.
†*Takifugu* annotations are from Ensembl version 23.2.1.
‡*Takifugu* annotations from Ensembl version 18.2.1 do not include UTRs.

tinuity at the level of N50 ultracontig size than the *Takifugu* draft sequence (7.62 Mb versus 125 kb). Critically, the anchoring of the assembly provides a comprehensive view of a fish genome sequence organized in individual chromosomes.

Genome landscape

A consequence of the remarkably compact nature of the *Tetraodon* genome is that its G+C content is much higher than in the larger genomes of mammals. Although the G+C content is shifted markedly, it still shows the same asymmetric bell-shaped distribution with an excess of higher values as seen in human and mouse (Fig. 2a). (G+C)-rich regions tend to be gene-rich in mammals, and analysis of our data shows that this is also true for *Tetraodon* (Fig. 2b, c). The *Tetraodon* genome thus cannot be considered as a single homogeneous component but, as in mammals, it is a mosaic of relatively gene-rich and gene-poor regions.

Transposable elements are very rare in the *Tetraodon* genome[12,13]: we estimate here that they do not exceed 4,000 copies; however, with 73 different types, they are richly represented (Supplementary Notes and Supplementary Table SI3). In sharp contrast, the human and mouse genomes contain only ~20 different types but are riddled with millions of transposable element copies. One of the intriguing features of the human genome is that the distribution of short interspersed nucleotide elements (SINEs) is biased towards (G+C)-rich regions, whereas long interspersed nucleotide elements (LINEs) favour (A+T)-rich regions. In *Tetraodon*, these preferences are precisely reverse: LINEs occur preferentially in (G+C)-rich regions and SINEs in (A+T)-rich regions (Fig. 2d). The reason for these differences is not clear.

The *Tetraodon* genome shows certain striking differences from the previously reported *Takifugu* genome sequence. *Takifugu* contains eightfold more copies of transposable elements[4] than *Tetraodon*, which may contribute to its slightly larger genome size (approximately 370 Mb; see Supplementary Information). More surprisingly, the G+C content of *Takifugu* does not show the characteristic asymmetry seen in mammals and in *Tetraodon* (Fig. 2a) nor the biases in SINE and LINE distribution (Supplementary Fig. S4). Why would the (G+C)-rich component be lacking in the *Takifugu* sequence, when this fraction is gene dense in mammals and in *Tetraodon*? This cannot be ascribed to transposable elements, which represent less than 5% of the assembly in both of these puffer fish species. One possible explanation is that the (G+C)-rich fraction exists in *Takifugu*, but was markedly under-represented as a result of aspects of the cloning, sequencing or assembly process. The fact that *Tetraodon* (G+C)-rich regions contain an excess of genes with no apparent orthologues in the *Takifugu* genome supports this hypothesis. Indeed, the *Tetraodon* genome appears to contain ~16.5% more coding exons than *Takifugu* (see below).

Tetraodon genes

Gene catalogue

The most prevalent features of the *Tetraodon* genome are protein-coding genes, which span 40% of the assembly. We constructed a catalogue of genes by adapting the GAZE[14] computational framework (Supplementary Fig. S5) in order to combine three types of data: *Tetraodon* complementary DNA mapping, similarities to human, mouse and *Takifugu* proteins and genomes, and *ab initio* gene models (Supplementary Notes and Supplementary Tables SI4 and SI5).

The current *Tetraodon* catalogue is composed of 27,918 gene models, with 6.9 coding exons per gene on average (7.3 including untranslated regions (UTRs); Table 2). Assuming that fish and mammal genes possess similar gene structures, this suggests that some *Tetraodon* annotated genes are partial or fragmented because human and mouse genes respectively show 8.7 and 8.4 coding exons per gene[2]. Adjusting the gene count for such fragmentation (by multiplying by 6.9/8.6) would yield an estimated gene count of 22,400 genes, whereas accounting for unsequenced regions of the genome might increase the estimate slightly further. Although such

Table 3 **Comparative InterPro analysis of fish, mammal and urochordate proteomes**

Tetraodon	*Takifugu*	Human	Mouse	*Ciona*	InterPro description
Actinopterygian-enriched					
61	78	22	21	48	Sodium:neurotransmitter symporter
33	29	11	13	33	Na^+/solute symporter
21	16	8	7	6	Sodium/calcium exchanger membrane region
141	191	86	97	52	Collagen triple helix repeat
15	28	6	4	19	HAT dimerization
17	15	5	4	27	Peptidase M12A, astacin
3	4	0	0	1	Inosine/uridine-preferring nucleoside hydrolase
Sarcopterygian-enriched					
0	0	275	173	0	KRAB box
0	0	14	8	0	KRAB-related
3	0	25	29	0	High mobility group protein HMG14 and HMG17
0	0	9	95	0	Vomeronasal receptor, type 1
0	0	13	21	0	Keratin, high sulphur B2 protein
0	0	3	3	0	Keratin, high-sulphur matrix protein
0	0	22	11	0	Mammalian taste receptor
0	0	11	9	0	Pancreatic RNase
0	0	7	8	0	β-Defensin
Vertebrate-enriched					
52	40	82	102	9	Histone core
252	253	240	228	88	Homeobox
62	56	80	55	9	Zn finger, B box
94	83	75	74	19	Zn-binding protein, LIM
65	56	70	135	17	HMG1/2 (high mobility group) box

Supplementary Table SI7 contains the top 100 InterPro domains in *Tetraodon*.

Table 4 **Evolutionarily conserved regions between mammals and fish**

Query genome	Target genome			
	Tetraodon nigroviridis	*Takifugu rubripes*	*Homo sapiens*	*Mus musculus*
Tetraodon nigroviridis	NA	ND	139,316	133,091
Takifugu rubripes	ND	NA	139,932	131,835
Combined fish	NA	NA	151,708	142,804
Homo sapiens	142,820	133,239	NA	ND
Mus musculus	140,407	129,996	ND	NA
Combined mammals	151,668	140,965	NA	NA

NA, not applicable; ND, not determined.

estimates are somewhat imprecise, it seems likely that *Tetraodon* has between 20,000–25,000 protein coding genes.

The *Tetraodon* gene catalogue appears to be the most complete so far for a fish, with coding exons and UTRs totalling ~36 Mb (~11% of the genome; Table 2). The *Takifugu* paper[4] reported an estimate of 35,180 genes, but it did not account for a high degree of fragmentation (~4.3 exons per gene model). More recent, unpublished analyses have revised this number sharply downward (Table 2). The human and *Tetraodon* genomes have a similar distribution of exon sizes but markedly different distributions of intron size (Supplementary Fig. S6a). Although neither genome seems to tolerate introns below approximately 50–60 base pairs, *Tetraodon* has accumulated a much higher frequency of introns at this lower limit. Interestingly, this phenomenon is not uniform across the genome: there is an excess of genes with many small introns (Supplementary Fig. S6b), suggesting that intron sizes fluctuate in a regional fashion.

Proteome comparison between vertebrates

We examined in detail two gene families with unusual properties that represent challenges for automatic annotation procedures and have particular biological interest. The first is the family of selenoproteins, where the UGA codon encodes a rare cysteine analogue named selenocysteine (Sec) instead of signalling the end of translation as in all other genes[15]. We annotated 18 distinct families in *Tetraodon* based on similarities with the 19 protein families known in eukaryotes, and discovered a new selenoprotein that seems to be restricted to the actinopterygians among vertebrates and does not have a Cys counterpart in mammals. We also catalogued type I helical cytokines and their receptors (HCRI), a group of genes that were not found in the *Takifugu* genome[4] because of their poor sequence conservation, leading to the hypothesis that fish may not possess this large family that includes hormones and interleukins. *Tetraodon,* in fact, contains 30 genes encoding HCRIs with a typical D200 domain (Supplementary Fig. S7) and represents all families previously described in mammals[16].

InterPro[17] domains were annotated in protein sequences predicted in the *Tetraodon*, *Takifugu*, human, mouse and the urochordate *Ciona intestinalis*[18] genome using InterProScan[19]. We did not identify major differences between fish and mammal InterPro families, except for a few striking cases (Table 3): (1) collagen molecules are much more diverse in fish than in mammals, with one *Tetraodon* gene containing 20 von Willebrand type A domains, the largest number found so far in a single protein. (2) Some domains associated with sodium transport are noticeably enriched in fishes and *Ciona*, perhaps a reflection of their adaptation to saline aquatic environments that was lost in land vertebrates. (3) Purine nucleosidases usually involved in the recovery of purine nucleosides are more abundant in fish, including an allantoin pathway for purine degradation that is present in *Tetraodon* and absent in human. (4) Several hundred KRAB box transcriptional repressors involved in chromatin-mediated gene regulation exist in mammals and are totally absent in fish. (5) Proteins involved in general gene regulation are more abundant in vertebrates than in *Ciona*.

Protein annotation with gene ontology (GO) classifications[20] shows only subtle differences between fish and mammals, as was already observed between human and mouse[2]. The largest differences between species are seen with the GO classification in molecular functions (Supplementary Fig. S9). Interestingly, the two puffer fish and *Ciona* often vary together, showing for instance a higher frequency of enzymatic and transporter functions, and a lower frequency of signal transducer and structural molecules than both mammals (human and mouse). These global observations are difficult to relate to evolutionary or physiological mechanisms but provide a framework to understand the emergence or decline of molecular functions in vertebrates.

Number of genes in mammals and teleosts

The total amount of coding sequence conserved between the two fish and the two mammalian genomes provides a measure of their respective coding capacity. The Exofish method[6] is well suited to measure this, because it translates entire genomes in all six frames and identifies conserved coding regions (ecores) with a high specificity and independently of prior genome annotation (Table 4; see also Supplementary Information). The four vertebrate genomes contain remarkably similar numbers of ecores, apart from minor differences attributable to varying degrees of sequence completion. This suggests that they possess fairly similar numbers of genes. In fact, the gene count may be slightly less in mammals than in fish because the proportion of ecores corresponding to pseudogenes is higher in mammals[21].

The human ecores can be used to search for previously unrecognized human genes. The discovery of new human genes is becoming an increasingly rare event, given the scale and intensity of international efforts to annotate the genome by systematic annotation pipelines and by human experts. Roughly 14,500 human ecores

Table 5 **Rates of DNA evolution in vertebrates**

Species	Total number of orthologues	Number of orthologues used	Average per cent identity (without gaps)	Observed number of substitutions per 4D site	Estimated amount of neutral evolution	Estimated rate of neutral evolution (sites per Myr)	K_a
Human–mouse	14,889	5,802	91.76	0.32	0.43	0.0057	0.05
Tetraodon–Takifugu	12,909	5,802	90.51	0.27	0.35	0.0146	0.06
Tetraodon–human	9,975	5,802	69.90	0.63	1.54*	–	0.24
Tetraodon–mouse	9,666	5,802	69.46	0.63	1.53*	–	0.25
Takifugu–human	9,143	5,802	70.05	0.63	1.52*	–	0.24
Takifugu–mouse	8,956	5,802	69.67	0.63	1.52*	–	0.25

*These values are saturated and cannot be considered reliable estimates.

conserved with *Tetraodon* sequences do not overlap any 'known' features (genes or pseudogenes) in the human genome. Using these as anchors for local gene identification using the GAZE program, we identified 904 novel human gene predictions. Of these, 63% are also supported by expressed sequence tag (EST) data (from human or other species) and 50% contain predicted InterPro protein domains (Supplementary Table SI9). The most convincing evidence supporting these gene predictions is that they are strongly enriched on chromosomes that have not yet been annotated by human experts (Supplementary Table SI10). The novel gene predictions have relatively small size (average coding sequence (CDS) of 469 bp), which may have caused them to be eliminated by systematic annotation procedures. They provide a rich resource to help complete the human gene catalogue.

Tetraodon gene evolution

We measured rates of sequence divergence between fish and mammals to estimate the relative speed with which functional and non-functional sequences evolve in these lineages. We used fourfold degenerate (4D) site substitutions in orthologous proteins as a proxy for neutral nucleotide mutations, an approach that has been shown to be robust across entire genomes[2]. To optimize further the selection of sites used for comparison, we only considered the 5,802 proteins that are identified as orthologues in all pairwise comparisons between human, mouse, *Tetraodon* and *Takifugu*. The average neutral nucleotide substitution rate, inferred using the REV model[22,23], shows that the divergence between *Tetraodon* and *Takifugu* is about twice as fast per year as between human and mouse (Table 5), or between mouse and rat[3].

We were interested to see whether this higher mutation rate is also seen in protein sequences. Pairwise comparison of all possible combinations of the 5,802 four-way orthologous proteins clearly indicates that proteins between the two puffer fish are more divergent than between the two mammals, despite the shorter evolutionary time that has elapsed (Fig. 3). This is confirmed by the fact that the average frequency of non-synonymous mutations (leading to an amino acid change, K_a) between *C. intestinalis* and human proteins is lower than between *Ciona* and *Tetraodon* (see Methods).

Independent of the overall rate of change, the ratio of non-synonymous to synonymous changes (K_a/K_s ratio) is much higher between the two puffer fish than between human and mouse (Supplementary Table SI11 and Supplementary Information), suggesting that protein evolution is proceeding more rapidly along the puffer fish lineage. The reasons for this faster tempo of protein change are unknown, although it is likely to be positively correlated with the higher rate of neutral mutation.

Genome evolution

Genome-wide sequence provides a rare opportunity to address key evolutionary questions in a global fashion, circumventing biases due to small sequence and gene samples. In this respect, the combination of long-range linkage in the *Tetraodon* sequence and its evolutionary divergence from the mammalian lineage at 450 Myr ago makes it possible to explore overall genome evolution in the vertebrate clade.

Evidence for whole-genome duplication

The occurrence of WGD in the ray-finned fish lineage is a hotly debated question due both to the cataclysmic nature of such an event and to the difficulty in establishing that it actually occurred[24–26].

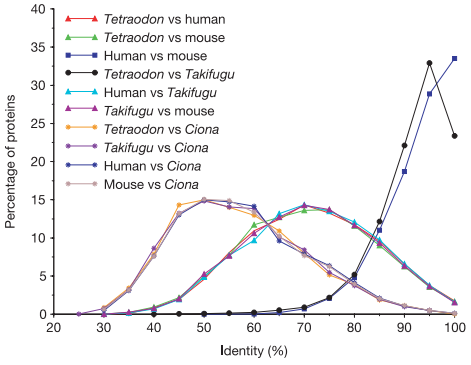


Figure 3 Distribution of the per cent identity between pairs of orthologous protein sets. Comparisons were performed with 2,289 proteins that are orthologous between the chordate *C. intestinalis* and all four vertebrates—*Tetraodon*, *Takifugu*, human and mouse (asterisks)—and with 5,802 proteins orthologous between all four vertebrates only, between fish and mammals (triangles) or between the two fish (circles), and between the two mammals (squares). As expected, all vertebrates show the same distribution profile compared to *Ciona* and both fish show the same distribution profile compared to mammals. Surprisingly, the distribution profile of the comparison between the two fish and between the two mammals is also very similar, despite the much shorter evolutionary time since the tetraodontiform radiation.

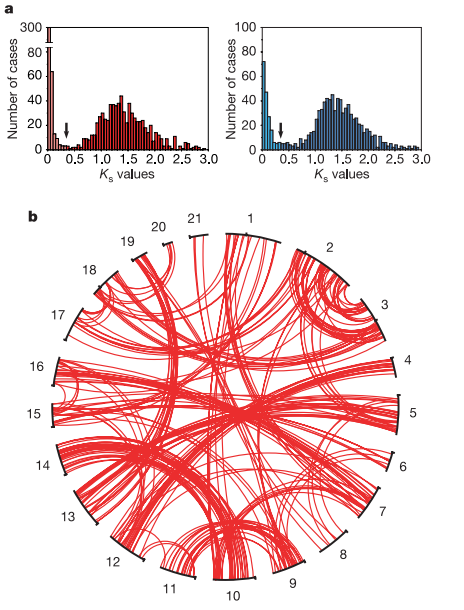


Figure 4 Genome duplication. **a**, Distribution of K_s values of duplicated genes in *Tetraodon* (left) and *Takifugu* (right) genomes. Duplicated genes broadly belong to two categories, depending on their K_s value being below or higher than 0.35 substitutions per site since the divergence between the two puffer fish (arrows). **b**, Global distribution of ancient duplicated genes ($K_s > 0.35$) in the *Tetraodon* genome. The 21 *Tetraodon* chromosomes are represented in a circle in numerical order and each line joins duplicated genes at their respective position on a given pair of chromosomes.

Definitive proof of WGD requires identifying certain distinctive signatures in long-range genome organization, which has previously been impossible to address with the data available.

It is expected that after WGD the resulting polyploid genome gradually returns to a diploid state through extensive gene deletion, with only a small proportion of duplicated copies ultimately retained as sources of functional innovation[26]. Paralogous chromosomes will thus each retain only a small subset of their initially common gene complement and then will be broken into smaller segments by genomic rearrangements. WGD will thus leave two distinctive signs for considerable periods before eventually fading.

The first distinctive sign is duplicated genes on paralogous chromosomes. In the absence of chromosomal rearrangement it would be simple to recognize two paralogous chromosomes arising from a WGD from the genome-wide distribution of duplicate genes: the chromosomes would each contain one member from many duplicated gene pairs occurring in the same order along their length. The difficulty is that this neat picture will eventually be blurred by interchromosomal rearrangement, which will disrupt the 1:1 correspondence between chromosomes, and intrachromosomal rearrangement, which will disrupt gene ordering along chromosomes.

We analysed the genome-wide distribution of duplicated gene pairs to see whether a strong correspondence between chromosomes could be detected. We identified 1,078 and 995 pairs of duplicated genes in the *Tetraodon* and *Takifugu* genomes, respectively, using conservative criteria (see Supplementary Information). On the basis of the frequencies of silent mutations (K_s) between copies, ~75% are 'ancient' duplications that arose before the *Tetraodon–Takifugu* speciation (Fig. 4a).

The chromosomal distribution of these ancient duplicates follows a striking pattern characteristic of a WGD. Genes on one chromosome segment have a strong tendency to possess duplicate copies on a single other chromosome (Fig. 4b). The correspondence is not a perfect 1:1 match owing to interchromosomal exchange, but it is vastly stronger than expected by chance (Supplementary Table SI12). As expected from a WGD, all chromosomes are involved. Remarkably, some duplicate chromosome pairs such as *Tetraodon* chromosome 9 (Tni9) and Tni11 have remained largely undisturbed by chromosome translocations since the duplication event. In other cases, one chromosome has links to two or three others, suggestive of either fusion or fragmentation (for example, Tni13 matches Tni5 and Tni19).

The second distinctive sign, which is an even more powerful signature of genome duplication, comes from comparison with a related species carrying a genome that did not undergo the WGD. Such a comparison was recently used to prove the existence of an ancient WGD in the yeast *Saccharomyces cerevisiae* based on comparison with a second yeast species *Kluyveromyces waltii* that diverged before the WGD[27,28]. Although two ancient paralogous regions typically retained only a few genes in common, they could be readily recognized because they showed a characteristic 2:1 mapping with interleaving; that is, they both showed conserved synteny and local order to the same region of the *K. waltii* genome with the *S. cerevisiae* genes interleaving in alternating stretches. Such regions were called blocks of DCS (doubly conserved synteny). Whereas the first distinctive sign of WGD depends only on a

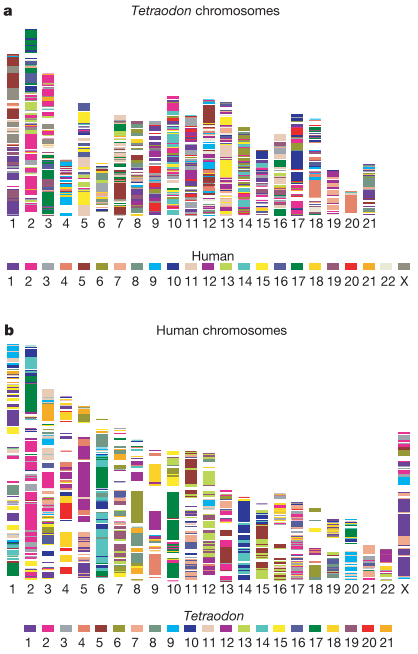


Figure 5 Synteny maps. **a**, For each *Tetraodon* chromosome, coloured segments represent conserved synteny with a particular human chromosome. Synteny is defined as groups of two or more *Tetraodon* genes that possess an orthologue on the same human chromosome, irrespective of orientation or order. *Tetraodon* chromosomes are not in descending order by size because of unequal sequence coverage. The entire map includes 5,518 orthologues in 900 syntenic segments. **b**, On the human genome the map is composed of 905 syntenic segments. See Supplementary Information for the synteny map between *Tetraodon* and mouse (Supplementary Fig. S11).

Table 6 **Distribution of human orthologues on *Tetraodon* chromosomes listed by their ancestral chromosome of origin**

	Ancestral chromosome											
	A	B	C	D	E	F	G	H	I	J	K	L
Tetraodon chromosome (copy 1)	4	17	2	2	5	13	7	1	1	10	9	6
Number of orthologues on copy 1	141	30	130	318	187	145	136	143	151	262	214	111
Percentage of orthologues on copy 1*	32.0	19.2	31.4	62.1	52.1	58.5	58.1	58.8	61.6	52.5	45.2	36.4
Tetraodon chromosome (copy 2)	12	18	3	3	13	19	16	7	15	14	11	8
Number of orthologues on copy 2	299	94	166	97	172	103	98	100	94	237	259	129
Percentage of orthologues on copy 2*	68.0	60.26	40.1	18.9	47.9	41.5	41.9	41.2	38.4	47.5	54.8	42.3
Tetraodon chromosome (copy 3)	–	20	18	17	–	–	–	–	–	–	–	21
Number of orthologues on copy 3	–	32	118	97	–	–	–	–	–	–	–	65
Percentage of orthologues on copy 3*	–	20.5	28.50	18.9	–	–	–	–	–	–	–	21.31

*Only orthologues that belong to syntenic groups are indicated here. For instance, ancestral chromosome A could be reconstructed with 141 *Tetraodon*–human orthologues belonging to *Tetraodon* chromosome 4 and 299 to chromosome 12.

minority of duplicated genes, the DCS signature considers all genes for which orthologues can be found in the related species.

We used 6,684 *Tetraodon* genes localized on individual chromosomes that possess an orthologue in either human or mouse to create a high-resolution synteny map (Fig. 5 and Supplementary Fig. S11, respectively). The map contains 900 syntenic groups composed of at least two consecutive genes (average 6.1; maximum 55) having orthologues on the same human chromosome; the syntenic groups include 76% of *Tetraodon*–human orthologues. The synteny map with mouse contains 1,011 syntenic groups, probably reflecting the higher degree of chromosomal rearrangement in the rodent lineage[2].

The synteny map typically associates two regions in *Tetraodon* with one region in human. Using precise criteria (see Methods) we defined DCS blocks for *Tetraodon* relative to human; in contrast to the yeast study, strict conservation of gene order within DCSs was not required. Notably, most (79.6%) orthologous genes in syntenic groups can be assigned to 90 DCS blocks (Fig. 6). As in *S. cerevisiae*[27], we see the distinctive interleaving pattern expected from WGD followed by massive gene loss. Analysis of the interleaving pattern shows that the gene loss occurred through many small deletions in a balanced fashion over the two *Tetraodon* sister chromosomes (average balance 42% and 58% of retention; Supplementary Information); this is consistent with the results in yeast.

These two analyses provide definitive evidence that the *Tetraodon* genome underwent a WGD sometime after its divergence from the mammalian lineage. The first test used only the ~3% of genes that represent duplicated gene pairs retained from the WGD. The second test used the pattern of 2:1 mapping with interleaving involving ~80% of orthologues between *Tetraodon* and human.

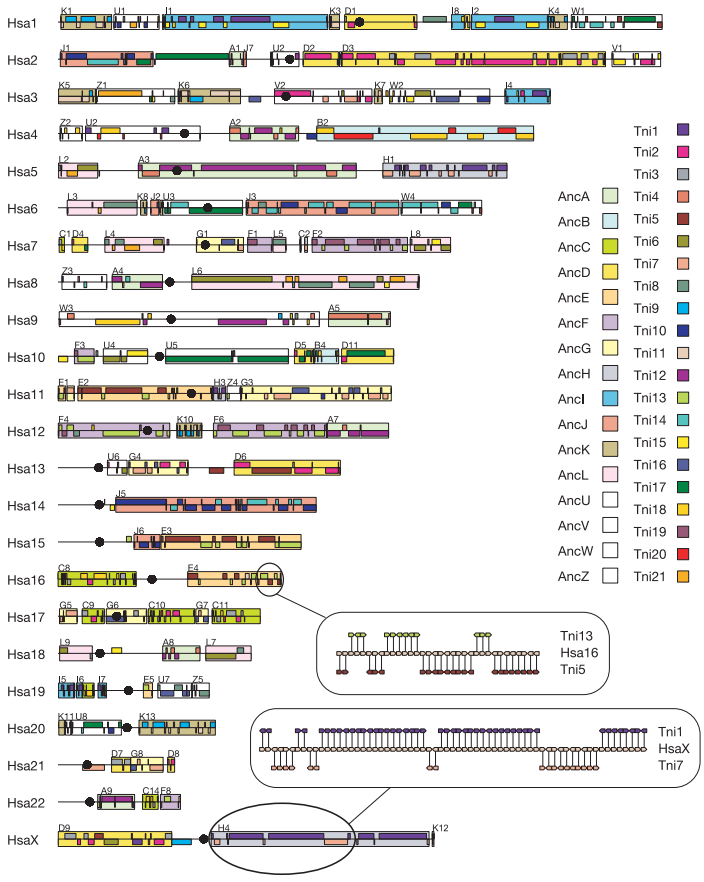


Figure 6 Duplicate mapping of human chromosomes reveals a whole-genome duplication in *Tetraodon*. Blocks of synteny along human chromosomes map to two (or three) *Tetraodon* chromosomes in an interleaving pattern. Small boxes represent groups of syntenic orthologous genes enclosed in larger boxes that define the boundaries of 110 DCS blocks. Black circles indicate human centromeres. A region of human chromosomes Xq and 16q are shown in detail with individual *Tetraodon* orthologous genes depicted on either side.

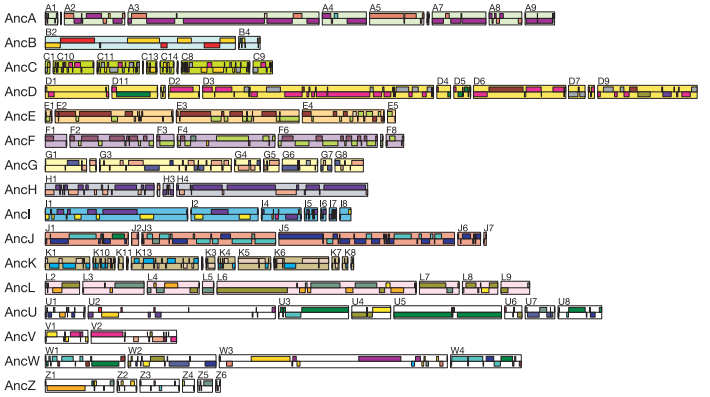


Figure 7 Composition of the ancestral osteichthyan genome. The 110 DCS blocks identified on the human genome are grouped according to their composition in terms of *Tetraodon* chromosomes, thus delineating 12 ancestral chromosomes containing 90 DCS blocks. The order of DCSs within an ancestral chromosome is arbitrary. The 20 blocks denoted by the letters U, V, W and Z (Supplementary Information) could not be assigned to an ancestral chromosome because each has a unique composition, probably due to rearrangements in the human or *Tetraodon* genome. Colour codes are as in Fig. 6.

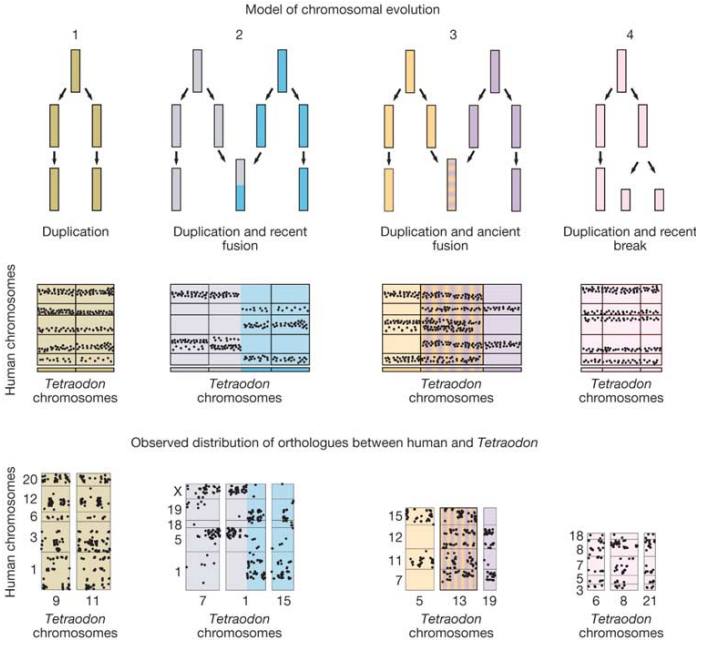


Figure 8 Reconstructing ancient genome rearrangements. Model of chromosome duplication followed by the four simplest chromosome rearrangements: (1) no rearrangement; (2) two different duplicate copies fused recently; (3) two different duplicate copies fused early after the duplication; (4) a duplicate chromosome fragmented very recently. In each model, the distribution of human orthologues from a given chromosomal region on two or three duplicate *Tetraodon* chromosomal regions is expected to be different (each dot is an orthologue, positioned in the human genome on the vertical axis and in the *Tetraodon* genome on the horizontal axis). The distinction between early or late events follows the assumption that intrachromosomal shuffling progressively redistributes genes over a given chromosome. A recent fusion would thus bring together two sets of genes that appear compartmented on their respective segments, whereas an ancient fusion shows the same pattern except that genes have been redistributed over the length of the fused chromosome. It should be noted that a fifth case exists, consisting of a chromosome break early after duplication but it is not represented here. The lower panel shows excerpts of data illustrating the four types of event. The complete Oxford grid is shown in Supplementary Fig. SI12.

The presence of supernumerary HOX clusters in zebrafish[7], *Tetraodon* (Fig. S8) and many other percomorphs[29] but not in the bichir *Polypterus senegalus*[30] indicates that the event has affected most teleosts but not all actinopterygians. This timing early in the teleost lineage is in agreement with recent evolutionary analyses in *Takifugu* that estimated the divergence time for most duplicated gene pairs at ~320–350 Myr ago[31,32].

The analyses above also shed light on the rate of intra- and interchromosomal exchange. The synteny analysis shows extensive syntenic segments in which gene content has been well preserved but gene order has been extensively scrambled (striking examples include conserved synteny of Tni20 with human chromosome 4q (Hsa4q) and Tni1 with HsaXq); this is consistent with observations in zebrafish[33]. The duplication analysis within *Tetraodon* also shows that the chromosomal correspondence of duplicated gene pairs has been extensively preserved, whereas local gene order has been largely scrambled. Both analyses thus indicate that a relatively high degree of intrachromosomal rearrangement and a relatively low degree of interchromosomal exchange have taken place in the *Tetraodon* lineage.

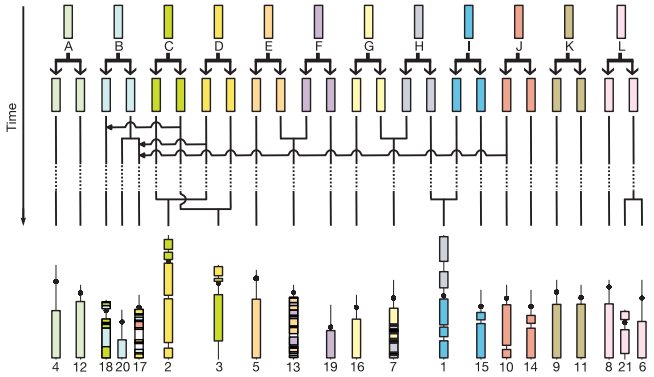


Figure 9 Model for the reconstruction of an ancestral bony vertebrate karyotype comprising 12 chromosomes, based on the pairing information provided by duplicated *Tetraodon* chromosomes showing interleaved patterns on human chromosomes. The ten major rearrangements (two ancient fusions, three recent fusions, one ancient and one recent fission, and three ancient translocations) are deduced by fitting the distribution of orthologues to the four simple theoretical models of chromosome evolution. The order between events is arbitrary although the approximate timeline differentiates between ancient and recent events respectively before and after the dashed line. Arrowheads point to the direction of three ancient translocations.

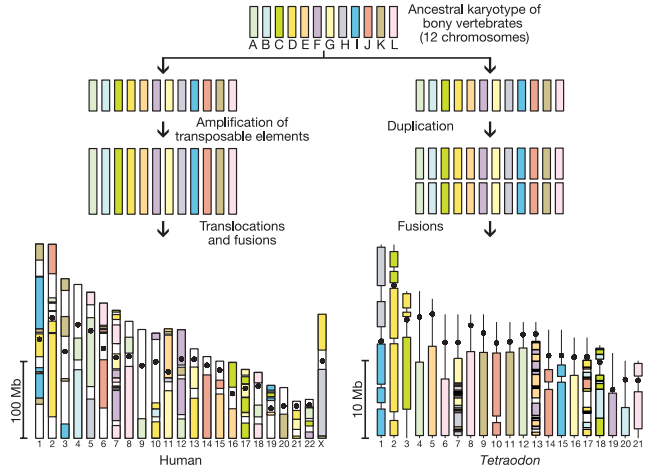


Figure 10 Proposed model for the distribution of ancestral chromosome segments in the human and the *Tetraodon* genomes. The composition of *Tetraodon* chromosomes is based on their duplication pattern (Fig. 9), whereas the composition of human chromosomes is based on the distribution of orthologues of *Tetraodon* genes (Fig. 6). A vertical line in *Tetraodon* chromosomes denotes regions where sequence has not yet been assigned. With 90 blocks in human compared with 44 in *Tetraodon*, the complexity of the mosaic of ancestral segments in human chromosomes underlines the higher frequency of rearrangements to which they were submitted during the same evolutionary period.

Ancestral genome of bony vertebrates

We then sought to use the correspondence between the *Tetraodon* and human genomes to attempt to reconstruct the karyotype of their osteichthyan (bony vertebrate) ancestor. The DCS blocks define *Tetraodon* regions that arose from duplication of a common ancestral region. Notably, the DCS blocks largely fall into 12 simple patterns: eight cases involving the interleaving of two current *Tetraodon* chromosomes and four cases involving three current *Tetraodon* chromosomes (Fig. 7 and Table 6). The first group represents cases in which the ancestral chromosomes have remained largely untouched by interchromosomal exchange; the second group represents cases in which one major translocation has occurred.

The distribution of *Tetraodon* orthologues in the human genome (shown as an Oxford grid in Supplementary Fig. S12) provides a detailed record that can be used to partially reconstruct the history of rearrangements in both lineages. We considered the expected distribution resulting from various types of interchromosomal rearrangements, assuming a relatively high degree of intrachromosomal shuffling (Fig. 8; see also Supplementary Information). We found that only ten large-scale interchromosomal events suffice to largely explain the data, connecting an ancestral vertebrate karyotype of 12 chromosomes to the modern *Tetraodon* genome of 21 chromosomes (Fig. 9). Eleven of the *Tetraodon* chromosomes appear to have undergone no major interchromosomal rearrangement. For example, 13 DCS blocks in human are composed of interleaved syntenic groups mapping to Tni9 and Tni11, which are presumed to be derived from a common ancestral chromosome denoted chromosome K (AncK; Fig. 7). The orthologue distribution between the two chromosomes (Fig. 8) confirms that they derive by duplication from AncK (Fig. 9). In a more complex case, Tni13 is systematically interleaved with Tni5 (AncE) or Tni19 (AncF), but Tni5 and Tni19 are never interleaved together; the orthologue distribution among the three chromosomes (Fig. 8) implies that the duplication partners of Tni5 and Tni19 fused soon after the WGD to give rise to Tni13 (Fig. 9). The overall model is consistent with a complete WGD, in that it accounts for all *Tetraodon* chromosomes.

Several lines of evidence support the historical reconstitution presented here. First, the pairing of *Tetraodon* chromosomes agrees with the independently derived distribution of duplicated genes in the genome (Fig. 4b). Second, centric fusions of the three largest chromosomes are consistent with cytogenetic studies[34], and the recent timing of the fusion leading to Tni1 is supported by cytogenetic studies showing its absence in *Takifugu*[35]. Third, the modal value for the haploid number of chromosomes in teleosts is 24 (refs 36–38), consistent with a WGD of an ancestral genome composed of 12 chromosomes.

The analysis also sheds light on genome evolution in the human lineage, with the interleaving patterns on human chromosomes delineating the mosaic of ancestral segments in the human genome (Figs 6 and 10). The results are consistent with and extend several known cases of rearrangements in the human lineage. The model correctly shows the recent fusion of two primate chromosomes leading to Hsa2 (ref. 39) occurring at the junction between two ancestral segments (D2 and D3; Fig. 6) in 2q13.2-2q14.1. It shows HsaXp and HsaXq to be of different origins (corresponding to AncD and AncH, respectively), consistent with the fact that HsaXp is known to be absent in non-placental mammals[40]. The map indicates that most of HsaXq and Hsa5q were once part of the same chromosome, but that the tip of HsaXq (Xq28) originates from a different ancestral segment and is thus a later addition. Some pairs of human chromosomes show similar or identical compositions, suggesting that they derived by fission from the same ancestral chromosome, with examples being Hsa13–Hsa21 and Hsa12–Hsa22; the latter case is consistent with cytogenetic studies showing that a fission occurred in the primate lineage[41].

The results show a major difference in the evolutionary forces shaping the *Tetraodon* and the human genomes (Fig. 10). Whereas 11 *Tetraodon* chromosomes did not undergo interchromosomal exchange over 450 Myr, only one human chromosome (Hsa14) was similarly undisturbed. Hsa7 is an extreme case, with contributions from six ancestral chromosomes. A possible explanation for the difference may be the massive integration of transposable elements in the human genome. The presence of transposable elements may increase the overall frequency of chromosome breaks, as well as the likelihood that a chromosome break fails to disrupt a gene (by increasing the size of intergenic intervals). It will be interesting to see whether teleosts that carry many more transposable elements (such as zebrafish) show a higher frequency of interchromosomal exchanges.

Conclusion

The purpose of sequencing the *Tetraodon* genome was to use comparative analysis to illuminate the human genome in particular and vertebrate genomes in general. The *Tetraodon* sequence, which has been made freely available during the course of this project, has already had a major impact on human gene annotation. It has provided the first clear evidence of a sharply lower human gene count[6] and has been used in the annotation of several human chromosomes[42–45]. Here, we show that it suggests an additional ~900 predicted genes in the human genome. Given its compact size, the *Tetraodon* genome will probably also prove valuable in identifying key conserved regulatory features in intergenic and intronic regions.

In addition, the *Tetraodon* genome provides fundamental insight into genome evolution in the vertebrate lineage. First, the analysis here shows that *Tetraodon* is the descendant of an ancient WGD that most probably affected all teleosts. Together with the recent demonstration of an ancient WGD in the yeast lineage, this suggests that WGD followed by massive gene loss may be an extremely important mechanism for eukaryote genome evolution—perhaps because it allows for the neofunctionalization of entire pathways rather than simply individual genes. There remains a fierce debate about whether one or more earlier WGD events occurred in early vertebrate evolution[25,46–50], with no direct and conclusive evidence found so far[51,52]. The examples of yeast and *Tetraodon* show that ultimate proof will probably best come from the sequence of a related non-duplicated species. An obvious candidate is amphioxus, as its non-duplicated status is supported by the presence of many single-copy genes (including one HOX cluster[53]) instead of two or more in vertebrates, and it is among our closest non-vertebrate relatives based on anatomical and evolutionary observations.

Second, the remarkable preservation of the *Tetraodon* genome after WGD makes it possible to infer the history of vertebrate chromosome evolution. The model suggests that the ancestral vertebrate genome was comprised of 12 chromosomes, was compact, and contained not significantly fewer genes than modern vertebrates (inasmuch as the WGD and subsequent massive gene loss resulted in only a tiny fraction of duplicate genes being retained). The explosion of transposable elements in the mammalian lineage, subsequent to divergence from the teleost lineage, may have provided the conditions for increased interchromosomal rearrangements in mammals; in contrast, the *Tetraodon* genome underwent much less interchromosomal rearrangement.

With the availability of additional vertebrate genomes (dog, marsupial, chicken, medaka, zebrafish and frog are underway), it will be possible to explore intermediate nodes such as the last common ancestor of amniotes, of sarcopterygians and of actinopterygians, and to gain an increasingly clearer picture of the early vertebrate ancestor. Because the early vertebrate genome is 'closer' to current invertebrates, this should in turn facilitate comparison between vertebrate and invertebrate evolution. □

articles

Methods

Sequencing, assembly and data access

Sequencing was performed as described previously for Genoscope[54] and the Broad Institute[1,2]. Approximately 4.2 million plasmid reads were cloned and sequenced from DNA extracted from two wild *Tetraodon* fish and passed extensive checks for quality and source, representing approximately 8.3-fold sequence coverage of the *Tetraodon* genome. To alleviate problems due to polymorphism, the assembly proceeded in four stages: (1) reads from a single fish were assembled by Arachne as described previously[10,11]; (2) reads from the second individual were added to increase sequencing depth; (3) scaffolds were constructed using plasmid and BAC paired reads; and (4) contigs from a separate assembly combining both individuals were added if they did not overlap with the first assembly. The final assembly can be downloaded from the EMBL/GenBank/DDBJ databases under accession number CAAE01000000. Full-length *Tetraodon* cDNAs have been submitted under accession numbers CR631133–CR735083. Ultracontigs organized in chromosomes are available from http://www.genoscope.org/tetraodon. This site also contains an annotation browser and further information on the project.

Gene annotation

Protein-coding genes were predicted by combining three types of information: alignments with proteins and genomic DNA from other species, *Tetraodon* cDNAs, and *ab initio* models. All alignments with genomic DNA from human and mouse were performed with Exofish as described previously[6], whereas a new Exofish method was developed to align *Takifugu* genomic DNA. Proteins predicted from human and mouse were also matched using Exofish and a selected subset was then aligned using Genewise. The integration of these data sources was performed with GAZE[14]. A specific GAZE automaton was designed, and parameters were adjusted on a training set of 184 manually annotated *Tetraodon* genes. See Supplementary Information for details.

Evolution of coding and non-coding DNA

To identify orthologous genes between human, mouse, *Tetraodon*, *Takifugu* and *Ciona*, their predicted proteomes were compared using the Smith–Waterman algorithm and reciprocal best matches were considered as orthologous genes between two species. However, only those genes that were reciprocal best matches between four or five species, and only sites that were aligned between the four or five genes, were further considered to compute the percentage identity, K_a, K_s and fourfold degenerate sites by the PBL method applying Kimura's two-parameter model[55–57]. See Supplementary Information for details.

Genome duplication

A core set of *Tetraodon* duplicated genes was identified by an all-against-all comparison of *Tetraodon* predicted protein using Exofish. Only proteins that matched a single other protein by reciprocal best match were considered further and realigned by the Smith–Waterman algorithm to compute K_a and K_s values. Duplicates with a $K_s > 0.35$ (the amount of neutral substitution since the *Tetraodon–Takifugu* divergence) were considered 'ancient' and used to calculate *P*-values for chromosome pairing (Supplementary Table SI12). Rules for classifying alternating patterns of syntenic groups along human chromosomes in DCS blocks included the following criteria: number of genes in syntenic groups, number of syntenic groups in the DCS region, number of *Tetraodon* chromosomes that alternate, and number of times the same combination of *Tetraodon* chromosomes occur in the human genome. See Supplementary Information for details.

Ancestral genome reconstruction

One category of DCS with the following definition encompassed most orthologues: "alternating series of i syntenic groups that belong to two ($i >= 2$) or three ($i >= 3$) *Tetraodon* chromosomes. The series may only be interrupted by groups from categories 'unassigned singletons' or 'background singletons'. A given combination of two or three *Tetraodon* chromosomes must appear at least twice in the human genome". These DCS blocks showed 12 recurring combinations of *Tetraodon* chromosomes, and were thus further classified in 12 groups labelled A to L. Each of the 12 groups, consisting of at least two DCS blocks with the same combination of alternating *Tetraodon* chromosomes, represents a proto-chromosome from the ancestral bony vertebrate (Osteichthyes). A model was then designed to account for the possible fates of chromosomes after duplication of the ancestral genome in the teleost lineage (Fig. 8). The model only deals with orthologous gene distribution between two genomes. It is simply based on the postulate that interchromosomal shuffling of genes within a genome increases with time, which is a measure to distinguish between ancient and recent events (for example, chromosome fusions or fissions). The two-dimensional distribution of 7,903 *Tetraodon*–human orthologues (Oxford Grid, Supplementary Fig. S12) was then confronted to the model and all 21 *Tetraodon* chromosomes could be grouped in pairs or triplets and assigned to a given type of event. See Supplementary Information for details.

Received 14 July; accepted 8 September 2004; doi:10.1038/nature03025.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
3. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
4. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).
5. Hedges, S. B. The origin and evolution of model organisms. *Nature Rev. Genet.* **3,** 838–849 (2002).
6. Roest Crollius, H. *et al.* Human gene number estimate provided by genome wide analysis using *Tetraodon nigroviridis* genomic DNA. *Nature Genet.* **25,** 235–238 (2000).
7. Amores, A. *et al.* Zebrafish hox clusters and vertebrate genome evolution. *Science* **282,** 1711–1714 (1998).
8. Robinson-Rechavi, M., Marchand, O., Escriva, H. & Laudet, V. An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. *Curr. Biol.* **11,** R458–R459 (2001).
9. Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. & Van de Peer, Y. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.* **13,** 382–390 (2003).
10. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12,** 177–189 (2002).
11. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).
12. Roest Crollius, H. *et al.* Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.* **10,** 939–949 (2000).
13. Bouneau, L. *et al.* An active non-LTR retrotransposon with tandem structure in the compact genome of the pufferfish *Tetraodon nigroviridis*. *Genome Res.* **13,** 1686–1695 (2003).
14. Howe, K. L., Chothia, T. & Durbin, R. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res.* **12,** 1418–1427 (2002).
15. Hatfield, D. L. *Selenium: Its Molecular Biology and Role in Human Health* (Kluwer, Dordrecht, 2001).
16. Boulay, J. L., O'Shea, J. J. & Paul, W. E. Molecular phylogeny within type I cytokines and their cognate receptors. *Immunity* **19,** 159–163 (2003).
17. Mulder, N. J. *et al.* InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief. Bioinform.* **3,** 225–235 (2002).
18. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298,** 2157–2167 (2002).
19. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848 (2001).
20. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32** (Database issue), D258–D261 (2004).
21. Torrents, D., Suyama, M., Zdobnov, E. & Bork, P. A genome-wide survey of human pseudogenes. *Genome Res.* **13,** 2559–2567 (2003).
22. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17,** 57–86 (1986).
23. Gu, X. & Li, W. H. A general additive distance with time-reversibility and rate variation among nucleotide sites. *Proc. Natl Acad. Sci. USA* **93,** 4671–4676 (1996).
24. Holland, P. W. H. Introduction: gene duplication in development and evolution. *Semin. Cell Dev. Biol.* **10,** 515–516 (1999).
25. Martin, A. Is tetralogy true? Lack of support for the "one-to-four" rule. *Mol. Biol. Evol.* **18,** 89–93 (2001).
26. Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.* **2,** 333–341 (2001).
27. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428,** 617–624 (2004).
28. Dietrich, F. S. *et al.* The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304,** 304–307 (2004).
29. Prohaska, S. J. & Stadler, P. F. The duplication of the Hox gene clusters in teleost fishes. *Theor. Biosci.* **123,** 89–110 (2004).
30. Chiu, C. H. *et al.* Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res.* **14,** 11–17 (2004).
31. Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A. & Van de Peer, Y. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl Acad. Sci. USA* **101,** 1638–1643 (2004).
32. Christoffels, A. *et al.* Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol.* **21,** 1146–1151 (2004).
33. Woods, I. G. *et al.* A comparative map of the zebrafish genome. *Genome Res.* **10,** 1903–1914 (2000).
34. Fischer, C. *et al.* Karyotype and chromosomal localization of characteristic tandem repeats in the pufferfish *Tetraodon nigroviridis*. *Cytogenet. Cell Genet.* **88,** 50–55 (2000).
35. Grutzner, F. *et al.* Classical and molecular cytogenetics of the pufferfish *Tetraodon nigroviridis*. *Chromosome Res.* **7,** 655–662 (1999).
36. Ohno, S., Wolf, U. & Atkin, N. B. Evolution from fish to mammals by gene duplication. *Hereditas* **59,** 169–187 (1968).
37. Ojima, Y. in *Chromosomes in Evolution of Eukaryotic Groups* (eds Sharma, A. K. & Sharma, A.) 111–145 (CRC Press, Boca Raton, 1983).
38. Naruse, K. *et al.* A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res.* **14,** 820–828 (2004).
39. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215,** 1525–1530 (1982).
40. Graves, J. A., Gecz, J. & Hameister, H. Evolution of the human X—a smart and sexy chromosome that controls speciation and development. *Cytogenet. Genome Res.* **99,** 141–145 (2002).
41. Richard, F., Lombard, M. & Dutrillaux, B. Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res.* **11,** 605–618 (2003).
42. The chromosome 21 mapping and sequencing consortium, The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
43. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).
44. Collins, J. E. *et al.* Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13,** 27–36 (2003).
45. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421,** 601–607 (2003).
46. Holland, P. W., Garcia-Fernandez, J., Williams, N. A. & Sidow, A. Gene duplications and the

origins of vertebrate development. *Development* (suppl.), 125–133 (1994).
47. Spring, J. Vertebrate evolution by interspecific hybridisation–are we polyploid? *FEBS Lett.* **400,** 2–8 (1997).
48. Friedman, R. & Hughes, A. L. Pattern and timing of gene duplication in animal genomes. *Genome Res.* **11,** 1842–1847 (2001).
49. Hughes, A. L., da Silva, J. & Friedman, R. Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res.* **11,** 771–780 (2001).
50. Thornton, J. W. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc. Natl Acad. Sci. USA* **98,** 5671–5676 (2001).
51. McLysaght, A., Hokamp, K. & Wolfe, K. H. Extensive genomic duplication during early chordate evolution. *Nature Genet.* **31,** 200–204 (2002).
52. Panopoulou, G. *et al.* New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res.* **13,** 1056–1066 (2003).
53. Garcia-Fernandez, J. & Holland, P. W. Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370,** 563–566 (1994).
54. Artiguenave, F. *et al.* Genomic exploration of the hemiascomycetous yeasts: 2. Data generation and processing. *FEBS Lett.* **487,** 13–16 (2000).
55. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–120 (1980).
56. Li, W. H., Wu, C. I. & Luo, C. C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2,** 150–174 (1985).
57. Pamilo, P. & Bianchi, N. O. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol. Biol. Evol.* **10,** 271–281 (1993).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements This work was supported by Consortium National de Recherche en Génomique. We thank T. Itami and S. Watabe for their gift of *Takifugu* blood samples; C. Nardon and M. Weiss for help with flow cytometry experiments; K. Howe for discussions regarding GAZE; R. Heilig for help with the annotation; the Centre Informatique National de l'Enseignement Supérieur for computer resources; and Gene-IT for assistance with the Biofacet software package.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.W. (jsbach@genoscope.cns.fr). The final assembly is available at EMBL/GenBank/DDBJ under accession number CAAE01000000. Full-length *Tetraodon* cDNAs have been deposited under accession numbers CR631133–CR735083; ultracontigs organized in chromosomes are available from http://www.genoscope.org/tetraodon.

Zebulon of the NeSL clade. Rex3 elements are widely spread in fish genomes[27]. We found nearly complete (but corrupted) copies of Rex3 and Babar in the *Tetraodon* genome, on the insert of fully sequenced BAC clones (C. Fischer, unpublished results). The Maui element, which is by far the most abundant retrotransposon of the Takifugu genome (and in which full-length copies are present), is far less abundant in the *Tetraodon* genome. Other distribution discrepancies between both pufferfish genomes can be seen with the I element, which is present only as fossils in Takifugu, while it is still moderately abundant in *Tetraodon*. Similar discrepancies in the abundance of transposons between these two genomes as well as among families (for instance, the TC1 family of *Tetraodon*), can also be observed.

5. *Tetraodon* Gene Annotation

5.1 Repeat Masking

Most of the genome comparisons were performed with repeat masked sequences. For this purpose, we searched and masked sequentially several kinds of repeats using BLASTN and TBLASTX:

- Microsatellites and known *Tetraodon* centromeric and subtelocentric satellite repeats
- *Tetraodon* specific transposable elements and rRNA sequences.
- Other eukaryotic known repeats and transposons available in Repbase.
- Tandem repeats with the TRF program[40].

5.2 Exofish between mammal genomes and *Tetraodon*

All Exofish comparisons between mouse or human and *Tetraodon* or *Takifugu* were performed using TBLASTX and filtering parameters described previously[41]. Computations were performed at the CINES (Centre Informatique National de l'Enseignement Supérieur) on a 768 CPU SGI ORIGIN 3800 computer, with the Biofacet software package from Gene-IT (www.gene-it.com).

5.3 Exofish between the *Takifugu* genome and *Tetraodon*

Compared to the Exofish version designed to detect ecores between mammals and fish, the much shorter evolutionary time that has elapsed since the divergence of the two pufferfish

imposed more severe alignment constraints, and required an additional filter in the form of Genscan[42] and GeneID[43] predictions. Both *ab initio* tools were trained on a set of manually annotated *Tetraodon* genes, and only alignments at least 60 bp long that overlapped a predicted exon from either program were considered valid. Calibrations were performed on two sets of *Tetraodon* gene. First, a set of 507 reference *Tetraodon* genes built using Genewise with human protein sequences, in which both the fish and the mammal genes have the same number of exons. The assumption is that this set of well conserved genes represents a stringent selection template to identify *Takifugu* alignments that must not occur in introns. The second set of reference genes comes from finished BAC sequences produced at Genoscope and elsewhere, in which 178 gene structures could be identified by human expertise using cDNAs sequences and comparisons with proteins from other species.

Comparisons between the *Tetraodon* and *Takifugu* genome assemblies using TBLASTN were much faster than between fish and mammals, and were computed on a cluster of 40 CPU alpha EV6.8 at Genoscope.

5.4 Genewise

In addition to gene structures provided by ecotigs, human and mouse proteins mapped to a given locus on the *Tetraodon* genome with Exofish were aligned using Genewise with default parameters[44]. In cases where several proteins from a given mammalian species overlapped on the *Tetraodon* assembly, the candidate with the longest span over the locus was chosen for a Genewise alignment.

5.5 cDNAs

An important resource for *Tetraodon* gene annotation came from cDNA sequences. They provide a high confidence evidence for the identification of protein coding genes, refine gene structures based on similar genes in other species and enable the detection of genes that evolve too fast for methods based on conservation during evolution.

We sequenced 286,955 cDNA clone ends corresponding to 155,067 clones from 7 libraries constructed from brain, muscle, liver, kidney, eye, ovaries and whole fish RNA preparations.

RNAs were extracted with the TRI REAGENT kit (Sigma) and polyA+ mRNAs were purified using a Quiagen Kit (Quiagen). Depending on the tissue, between 11 (muscle and liver) and 32 (kidneys and brain) fish were required to obtain sufficient material, i.e. between

50 and 500 µg of polyA+ mRNA, except for the whole fish where a single individual was sufficient.

Thirteen full-length cDNA libraries were constructed according to the method of Oligo-Capping as described by Maruyana and Sugano[45]. The Oligo-Capping method includes three enzymatic reaction steps. First, BAP (Bacterial alcaline phosphatase, Takara, 1.2U) hydrolyses the phosphate of truncated mRNA 5' ends whose cap structures have been truncated. Then, the tobacco acid pyrophosphatase (TAP, TEBU, 40 U) removes the cap structure leaving a free phosphate at the 5' end of complete mRNAs. Third, the T4 RNA ligase (Takara, 250U), which requires a phosphate at the 5'end as its substrate, selectively ligates 5' r-oligos which contains a SfiI site only to the 5' ends that originally had the cap structure. Using Oligo-capped mRNA, first-strand cDNA was synthetised with dT adapter primers by RNaseH reverse transcriptase (SuperScriptII RNase H Reverse Transcriptase, Invitrogen, 400 U). Finally, after alkaline degradation of the RNA, first-strand cDNAs were amplified by PCR (20 cycles at 94°C for 1min, 56°C for 1 min, 72°C for 10 min) using the LA Taq (Kit Takara, 10 U), and digested with restriction enzyme SfiI.

For cloning, inserts were ligated in a plasmid vector (pME18S-FL3; Maruyama and Sugano, accession AB009864) using the DNA Ligation kit Ver.1. (Takara). Ligations were electroporated in E.coli DH10B cells and plated on LB agar with ampicilin. After overnight growth, single colonies were robotically picked in 384 microtitre plates and frozen at -80°C. DNA extractions and sequencing was performed as previously described[4].

We did not perform quality checks on the cDNAs reads for two reasons. First, corrupted cDNA sequence reads (e.g. continuous run of a mononucleotide) may score high Phred values and second, we postulated that aligning cDNAs sequences to the assembled genomic DNA would select the useful reads from those that are of too low quality. To align the cDNAs we used BLAST against the microsatellite masked assembly with the following parameters: W=20, X=8, match=5, mismatch=-4. The scores of all High Scoring Pairs (HSPs) are then summed in each genomic interval where the cDNA end sequence matches, and the interval with the highest score is selected if it scores above 1,000. In cases where two intervals have equal scores, both are selected. The interval corresponding to the 5' and the 3' sequences of the same clone were then fused if they lied on the same scaffold and if they were separated by less than 30 kb. Only 91% of the 286,955 sequences could be aligned in this way. Those that did not match represent vector only clones (4%) and low quality sequences or genes absent from the assembly (5%). To estimate the fraction of cDNAs really missing from the assembly,

we aligned the 24,758 cDNA sequences that did not match *Tetraodon*, to the *Takifugu* assembly, and obtained only 650 positive hits corresponding to 136 clusters. Since it is unlikely that large unsequenced regions of both pufferfish genomes overlap extensively except in gene poor heterochromatin, we conclude that the vast majority of cDNA sequences that did not align to the *Tetraodon* assembly are low quality reads or contaminations. To create gene models, 5' and 3' cDNA sequences where first assembled by Phrap when possible (99,204 pairs), and aligned using EST_GENOME[46] to the genomic interval identified by BLAST that was first extended by 5 kb on each side. Parameters for EST_GENOME were: mismatch=2 and penalty=3. We obtained 147,835 gene models and from those, we eliminated 13,481 models that were considered unreliable: unspliced models and models overlapping on the forward and reverse strand. The remaining 134,354 models were individually provided to GAZE[47], and they represent 12,154 clusters on the assembly.

5.6 Geneid and Genscan

Geneid[48]and Genscan[49] *ab inito* gene prediction software were trained on 184 *Tetraodon* genes that had been annotated and reviewed in finished sequenced by human experts. We then identified the same genes in the genome assembly and reconstituted one long sequence from the 184 genes in draft sequence. The performance[50] of GeneID and Genscan on this sequence were respectively 46% and 41% for specificity, and respectively 59% and 49% for sensitivity.

5.7 Integration of resources using GAZE

All the resources described here were used to automatically build *Tetraodon* gene models using GAZE[47]. Individual predictions from each of the programs (GeneID, Genscan, Exofish, Genewise, EST_GENOME) were broken down into segments (coding, intron, intergenic) and signals (start codon, stop codon, splice acceptor, splice donor, transcript start, transcript stop). Segments and signals were extracted from Genewise and EST_GENOME alignments. Because geneid and Genscan exons are not specific, we only considered their signals (splice donor, splice acceptor, etc.) but did not use their exons as "coding segments". Ecores and ecotigs do not predict exon boundaries so they were only used to generate "coding segments" but no signals (Fig. S6B). Each segment or signal from a given program was given a value reflecting our confidence in the data, and these values were used as scores for the arcs of the GAZE automaton (Fig. S6A). All signals from a given source were given a fixed score, but

segment scores were context sensitive: coding segment scores were linked to the percentage identity (%ID) of the alignment; intronic segment scores were linked to the %ID of the flanking exons; the intergenic segment score was linked to the score of the flanking ecotigs. All scores were then homogenised on scale from 1 to 100. Finally, the impact of each data source (Exofish, geneid, etc.) was evaluated on a reference *Tetraodon* sequence containing 184 genes individually annotated by human experts, and a weight was assigned to each resource to further reflect its reliability and accuracy in predicting gene models. This weight acts as a multiplicator for the score of each information source, before processing by GAZE. On the reference sequence, the final selection of coding segments, signals, associated scores and weights, once processed by GAZE, generates models with 72% sensitivity and 74% specificity in exons.

When applied to the entire assembled sequence, GAZE predicts 34,355 gene models. We used a filter to reject most obvious artefacts (CDS with a single amino acid for instance). Criteria for rejections were: 1,210 models with a CDS smaller than 75 bp; 2,997 models with more than 50% of their exons with a GAZE score below 0 (i.e. exon only supported by ab initio methods); 1,000 models with an overall GAZE score below 1,000.

6. Analysis of specific gene families

6.1 Class I cytokines and their receptors

The sequences from the known class I helical cytokines (HC) and their receptors (HCR) from vertebrates were used to search both the *Tetraodon* predicted peptides and the translated assembly sequence. Matching *Tetraodon* sequences could be classified in three categories:

1) Protein sequences clearly belonging to the HC or HCR families
2) Already identified proteins not belonging to HC or HCR familes
3) Sequences that do not clearly belong to either category 1 or 2.

Genes models were built around initial sequence alignments from categories 1 and 3 and compared to canonical gene structures coding for helical cytokines (phase 0 introns) or for the D200 domain of their receptors (phase1,2,1,0,1 introns). Using this strategy, 9 genes potentially coding for class I helical cytokines were identified and 30 potentially coding for their receptors. For each gene, the most robust putative exons were chosen to design oligonucleotides that could be used for Q-RT-PCR to test for their expression and look for tissues with the highest expression. For each gene, RNAs from the tissue showing the highest

***Table S4*. Summary of evidence (coding segments) used to annotate the *Tetraodon* genome**

Type of ressource	Number of features (Predictions)		GAZE annotations supported by at least one feature of the resource	GAZE annotations exclusively supported with the ressource		GAZE annotations supported by the resource plus methods of type 1		GAZE annotations supported by the resource plus methods of type 2		GAZE annotations supported by the resource plus methods of type 3	
				Genes	**Exons**	**Genes**	**Exons**	**Genes**	**Exons**	**Genes**	**Exons**
1. ExoFish	**Ecotigs**	**Ecores**									
Exofish with human	29,748	142,849	22,727	n.a.	n.a.	-	-	47	117	5	72
Exofish with mouse	29,517	141,647	22,587	n.a.	n.a.	-	-	49	89	3	65
Exofish with Takifugu	17,776	192,352	25,228	n.a.	n.a.	-	-	2,410	11,274	25	91
Exofish with human IPI	27,900	176,455	23,228	n.a.	n.a.	-	-	0	351	23	1,077
Exofish with mouse IPI	27,123	165,741	22,384	n.a.	n.a.	-	-	1	165	4	444
2. *ab-initio* methods	**Genes**	**Exons**									
Genscan	28,059	199,234	24,577	0	2,693	843	5,337	-	-	48	3,427
GeneID	29,415	186,922	24,565	0	2,135	611	7,742	-	-	46	4,627
3. Est_genome and Genewise	**Genes**	**Exons**									
cDNAs	12,154†	n.d.	7,008	270	6,183	340	6,768	510	9,529	-	-
Genewise with human IPI	21,692	163,802	21,704	0	924	172	2,935	0	1,611	-	-
Genewise with mouse IPI	21,567	154,366	21,059	0	0	0	0	0	218	-	-

n.a., not applicable. The GAZE could not predict a gene with these annotation because they do not provide signals.

†: The cDNAs were first clustered on the basis of their overlap on genomic DNA to allow comparisons with other features in this table.

n.d., not determined. Because multiple exons of cDNAs may overlap over the same exons in genomic DNA, it did not make sense to sum the total number of cDNA exons for this table.

Table S5. **Summary of evidence (signals) used to annotate the *Tetraodon* genome**

Features		transcript_start	transcript_stop	start	stop	splice3	splice5
Number of features available		7,068	5,651	27,918	27,918	168,966	168,966
Real number of features used		5,425	5,644	25,412	17,830	168,966	168,966
cDNAs	available	42,110	37,161	526,700	n.a.	94,346	94,146
	used	5,945 14%	5,976 16%	5,126 (<1%)	n.a.	70,182 (75%)	70,358 (75%)
Genewise with human IPI	available	n.a.	n.a.	21,692	21,692	142,210	142,210
	used	n.a.	n.a.	13,691 (63%)	14,908 (69%)	121,649 (85%)	122,319 (86%)
Genewise with mouse IPI	available	n.a.	n.a.	21,567	21,567	132,799	132,799
	used	n.a.	n.a.	8,650 (40%)	10,271 (48%)	109,309 (82%)	110,384 (83%)
Genscan	available	n.a.	n.a.	18,972	20,086	180,262	179,148
	used	n.a.	n.a.	7,557 (40%)	3,500 (17%)	111,306 (62%)	112,709 (63%)
GeneID	available	n.a.	n.a.	17,893	19,791	169,029	167,131
	used	n.a.	n.a.	7,176 (40%)	3,265 (16%)	103,365 (61%)	103,376 (62%)
Annotations exclusively supported with the resource	cDNAs	5,425	5,644	2,615	n.a.	7,309	7,458
	Genewise with human IPI	n.a.	n.a.	4,715	4,461	5,487	4,928
	Genewise with mouse IPI	n.a.	n.a.	22	30	0	0
	Genscan	n.a.	n.a.	2,921	948	9,912	9,792
	GeneID	n.a.	n.a.	2,590	746	7,204	7,343

Initial sequencing and comparative analysis of the mouse genome

Mouse Genome Sequencing Consortium (including G. Parra and R. Guigó).
Nature 420(6915):520-562 (2002).

The mouse genome was the second mammalian genome sequenced, and it was a really genomic breakthrough as it provides the key to discover the secrets of our own DNA. It allows for the first time the complete comparative analysis of two mammalian genomes. In addition, the mouse genome encodes an experimentally tractable organism. This means that it is now possible to determine the function of each and every component gene by experimental manipulation and evaluation, in the context of the whole organism.

The two genomes, are remarkably similar: 99% of mouse genes seem to have a direct human counterpart. On the other hand, only about 40% of the complete genomic sequence can be aligned. Therefore, it means that most of the divergences between human and mouse seem to occur in the non-coding DNA regions.

Our contribution to this work was basically in the *De novo gene prediction* section, pages 539-540. This section give some insights of the number of missed genes in the first conservative mouse annotations. *De novo*, refers to the fact that the analyzed gene predictions are only based in comparative genomic methods, without using any homology based on proteins or expression evidences databases. Therefore, these predictions are supposed to be genes without a strong homology to any known protein.

The section analyzes the results of `sgp2` and `twinscan` emphasizing the similarities and differences against the ENSEMBL automatic annotation pipeline. Most of the comments and results are derived from the paper presented in chapter 6. Therefore, my main contribution was to filter `sgp2` and `twinscan` predictions and to obtain the corresponding statistics of the overlapping ratios against ENSEMBL.

Due to the length of the corresponding paper (42 pages), it what follows its only reproduced the first page of the paper and the two pages corresponding *De novo gene prediction* section, where the gene comparative prediction result are discussed.

articles

Initial sequencing and comparative analysis of the mouse genome

Mouse Genome Sequencing Consortium*

**A list of authors and their affiliations appears at the end of the paper*

The sequence of the mouse genome is a key informational tool for understanding the contents of the human genome and a key experimental tool for biomedical research. Here, we report the results of an international collaboration to produce a high-quality draft sequence of the mouse genome. We also present an initial comparative analysis of the mouse and human genomes, describing some of the insights that can be gleaned from the two sequences. We discuss topics including the analysis of the evolutionary forces shaping the size, structure and sequence of the genomes; the conservation of large-scale synteny across most of the genomes; the much lower extent of sequence orthology covering less than half of the genomes; the proportions of the genomes under selection; the number of protein-coding genes; the expansion of gene families related to reproduction and immunity; the evolution of proteins; and the identification of intraspecies polymorphism.

With the complete sequence of the human genome nearly in hand[1,2], the next challenge is to extract the extraordinary trove of information encoded within its roughly 3 billion nucleotides. This information includes the blueprints for all RNAs and proteins, the regulatory elements that ensure proper expression of all genes, the structural elements that govern chromosome function, and the records of our evolutionary history. Some of these features can be recognized easily in the human sequence, but many are subtle and difficult to discern. One of the most powerful general approaches for unlocking the secrets of the human genome is comparative genomics, and one of the most powerful starting points for comparison is the laboratory mouse, *Mus musculus*.

Metaphorically, comparative genomics allows one to read evolution's laboratory notebook. In the roughly 75 million years since the divergence of the human and mouse lineages, the process of evolution has altered their genome sequences and caused them to diverge by nearly one substitution for every two nucleotides (see below) as well as by deletion and insertion. The divergence rate is low enough that one can still align orthologous sequences, but high enough so that one can recognize many functionally important elements by their greater degree of conservation. Studies of small genomic regions have demonstrated the power of such cross-species conservation to identify putative genes or regulatory elements[3–12]. Genome-wide analysis of sequence conservation holds the prospect of systematically revealing such information for all genes. Genome-wide comparisons among organisms can also highlight key differences in the forces shaping their genomes, including differences in mutational and selective pressures[13,14].

Literally, comparative genomics allows one to link laboratory notebooks of clinical and basic researchers. With knowledge of both genomes, biomedical studies of human genes can be complemented by experimental manipulations of corresponding mouse genes to accelerate functional understanding. In this respect, the mouse is unsurpassed as a model system for probing mammalian biology and human disease[15,16]. Its unique advantages include a century of genetic studies, scores of inbred strains, hundreds of spontaneous mutations, practical techniques for random mutagenesis, and, importantly, directed engineering of the genome through transgenic, knockout and knockin techniques[17–22].

For these and other reasons, the Human Genome Project (HGP) recognized from its outset that the sequencing of the human genome needed to be followed as rapidly as possible by the sequencing of the mouse genome. In early 2001, the International Human Genome Sequencing Consortium reported a draft sequence covering about 90% of the euchromatic human genome, with about 35% in finished form[1]. Since then, progress towards a complete human sequence has proceeded swiftly, with approximately 98% of the genome now available in draft form and about 95% in finished form.

Here, we report the results of an international collaboration involving centres in the United States and the United Kingdom to produce a high-quality draft sequence of the mouse genome and a broad scientific network to analyse the data. The draft sequence was generated by assembling about sevenfold sequence coverage from female mice of the C57BL/6J strain (referred to below as B6). The assembly contains about 96% of the sequence of the euchromatic genome (excluding chromosome Y) in sequence contigs linked together into large units, usually larger than 50 megabases (Mb).

With the availability of a draft sequence of the mouse genome, we have undertaken an initial comparative analysis to examine the similarities and differences between the human and mouse genomes. Some of the important points are listed below.

- The mouse genome is about 14% smaller than the human genome (2.5 Gb compared with 2.9 Gb). The difference probably reflects a higher rate of deletion in the mouse lineage.
- Over 90% of the mouse and human genomes can be partitioned into corresponding regions of conserved synteny, reflecting segments in which the gene order in the most recent common ancestor has been conserved in both species.
- At the nucleotide level, approximately 40% of the human genome can be aligned to the mouse genome. These sequences seem to represent most of the orthologous sequences that remain in both lineages from the common ancestor, with the rest likely to have been deleted in one or both genomes.
- The neutral substitution rate has been roughly half a nucleotide substitution per site since the divergence of the species, with about twice as many of these substitutions having occurred in the mouse compared with the human lineage.
- By comparing the extent of genome-wide sequence conservation to the neutral rate, the proportion of small (50–100 bp) segments in the mammalian genome that is under (purifying) selection can be estimated to be about 5%. This proportion is much higher than can be explained by protein-coding sequences alone, implying that the genome contains many additional features (such as untranslated regions, regulatory elements, non-protein-coding genes, and chromosomal structural elements) under selection for biological function.
- The mammalian genome is evolving in a non-uniform manner,

comes from a single collection of mouse cDNAs (the initial RIKEN cDNAs[41]). These cDNAs are very short on average, with few exons (median 2) and small ORFs (average length of 85 amino acids); whereas some of these may be true genes, most seem unlikely to reflect true protein-coding genes, although they may correspond to RNA genes or other kinds of transcripts. Both groups were omitted in the comparative analysis below.

Comparison of mouse and human gene sets

We then sought to assess the extent of correspondence between the mouse and human gene sets. Approximately 99% of mouse genes have a homologue in the human genome. For 96% the homologue lies within a similar conserved syntenic interval in the human genome. For 80% of mouse genes, the best match in the human genome in turn has its best match against that same mouse gene in the conserved syntenic interval. These latter cases probably represent genes that have descended from the same common ancestral gene, termed here 1:1 orthologues.

Comprehensive identification of all orthologous gene relationships, however, is challenging. If a single ancestral gene gives rise to a gene family subsequent to the divergence of the species, the family members in each species are all orthologous to the corresponding gene or genes in the other species. Accordingly, orthology need not be a 1:1 relationship and can sometimes be difficult to discern from paralogy (see protein section below concerning lineage-specific gene family expansion).

There was no homologous predicted gene in human for less than 1% (118) of the predicted genes in mouse. In all these cases, the mouse gene prediction was supported by clear protein similarity in other organisms, but a corresponding homologue was not found in the human genome. The homologous genes may have been deleted in the human genome for these few cases, or they could represent the creation of new lineage-specific genes in the rodent lineage—this seems unlikely, because they show protein similarity to genes in other organisms. There are, however, several other possible reasons why this small set of mouse genes lack a human homologue. The gene predictions themselves or the evidence on which they are based may be incorrect. Genes that seem to be mouse-specific may correspond to human genes that are still missing owing to the incompleteness of the available human genome sequence. Alternatively, there may be true human homologues present in the available sequence, but the genes could be evolving rapidly in one or both lineages and thus be difficult to recognize. The answers should become clear as the human genome sequence is completed and other mammalian genomes are sequenced. In any case, the small number of possible mouse-specific genes demonstrates that *de novo* gene addition in the mouse lineage and gene deletion in the human lineage have not significantly altered the gene repertoire.

Mammalian gene count

To re-estimate the number of mammalian protein-coding genes, we studied the extent to which exons in the new set of mouse cDNAs sequenced by RIKEN[132] were already represented in the set of exons contained in our initial mouse gene catalogue, which did not use this set as evidence in gene prediction. This cDNA collection is a much broader and deeper survey of mammalian cDNAs than previously available, on the basis of sampling of diverse embryonic and adult tissues[150]. If the RIKEN cDNAs are assumed to represent a random sampling of mouse genes, the completeness of our exon catalogue can be estimated from the overlap with the RIKEN cDNAs. We recognize this assumption is not strictly valid but nonetheless is a reasonable starting point.

The initial mouse gene catalogue of 191,290 predicted exons included 79% of the exons revealed by the RIKEN set. This is an upper bound of sensitivity as some RIKEN cDNAs are probably less than full length and many tissues remain to be sampled. On the basis of the fraction of mouse exons with human counterparts, the percentage of true exons among all predicted exons or the specificity of the initial mouse gene catalogue is estimated to be 93%. Together, these estimates suggest a count of about 225,189 exons in protein-coding genes in mouse ($191{,}290 \times 0.93/0.79$).

To estimate the number of genes in the genome, we used an exon-level analysis because it is less sensitive to artefacts such as fragmentation and pseudogenes among the gene predictions. One can estimate the number of genes by dividing the estimated number of exons by a good estimate of the average number of exons per gene. A typical mouse RefSeq transcript contains 8.3 coding exons per gene, and alternative splicing adds a small number of exons per gene. The estimated gene count would then be about 27,000 with 8.3 exons per gene or about 25,000 with 9 exons per gene. If the sensitivity is only 70% (rather than 79%), the exon count rises to 254,142, yielding a range of 28,000–30,500.

In the next section, we show that gene predictions that avoid many of the biases of evidence-based gene prediction result in only a modest increase in the predicted gene count (in the range of about 1,000 genes). Together, these estimates suggest that the mammalian gene count may fall at the lower end of (or perhaps below) our previous prediction of 30,000–40,000 based on the human draft sequence[1]. Although small, single-exon genes may add further to the count, the total seems unlikely to greatly exceed 30,000. This lower estimate for the mammalian gene number is consistent with other recent extrapolations[141]. However, there are important caveats. It is possible that the genome contains many additional small, single-exon genes expressed at relatively low levels. Such genes would be hard to detect by our various techniques and would also decrease the average number of exons per gene used in the analysis above.

De novo gene prediction

The gene predictions above have the strength of being based on experimental evidence but the weakness of being unable to detect new exons without support from known transcripts or homology to known cDNAs or ESTs in some organism. In particular, genes that are expressed at very low levels or that are evolving very rapidly are less likely to be present in the catalogue (R. Guigó, unpublished data).

Ideally, one would like to perform *de novo* gene prediction directly from genomic sequence by recognizing statistical properties of coding regions, splice sites, introns and other gene features. Although this approach works relatively well for small genomes with a high proportion of coding sequence, it has much lower specificity when applied to mammalian genomes in which coding sequences are sparser. Even the best *de novo* gene prediction programs (such as GENSCAN[145]) predict many apparently false-positive exons.

In principle, *de novo* gene prediction can be improved by analysing aligned sequences from two related genomes to increase the signal-to-noise ratio[135]. Gene features (such as splice sites) that are conserved in both species can be given special credence, and partial gene models (such as pairs of adjacent exons) that fail to have counterparts in both species can be filtered out. Together, these techniques can increase sensitivity and specificity.

We developed three new computer programs for dual-genome *de novo* gene prediction: TWINSCAN[160,325], SGP2 (refs 161, 326) and SLAM[162]. We describe here results from the first two programs. The results of the SLAM analysis can be viewed at http://bio.math.berkeley.edu/slam/mouse/. To predict genes in the mouse genome, these two programs first find the highest-scoring local mouse–human alignment (if any) in the human genome. They then search for potential exonic features, modifying the probability scores for the features according to the presence and quality of these human alignments. We filtered the initial predictions of these programs, retaining only multi-exon gene predictions for which there were corresponding consecutive exons with an intron in an aligned position in both species[327].

After enrichment based on the presence of introns in aligned

locations, TWINSCAN identified 145,734 exons as being part of 17,271 multi-exon genes. Most of the gene predictions (about 94%) were present in the above evidence-based gene catalogue. Conversely, about 78% of the predicted genes and about 81% of the exons in this catalogue were at least partially represented by TWINSCAN predictions. TWINSCAN predicted an extra 4,558 (3%) new exons not predicted by the evidence-based methods. SGP2 produced qualitatively similar results. The total number of predicted exons was 168,492 contained in 18,056 multi-exon genes, with 86% of the predicted genes in the evidence-based gene catalogue at least partially represented. Approximately 83% of the exons in the catalogue were detected by SGP2, which predicted an additional 9,808 (6%) new exons. There is considerable overlap between the two sets of new predicted exons, with the TWINSCAN predictions largely being a subset of the SGP2 predictions; the union of the two sets contains 11,966 new exons.

We attempted to validate a sample of 214 of the new predictions by performing PCR with reverse transcription (RT) between consecutive exons using RNA from 12 adult mouse tissues[163] and verifying resulting PCR products by direct DNA sequencing. Our sampling involved selecting gene predictions without nearby evidence-based predictions on the same strand and with an intron of at least 1 kb. The validation rate was approximately 83% for TWINSCAN and about 44% for SGP2 (which had about twice as many new exons; see above). Extrapolating from these success rates, we estimate that the entire collection would yield about 788 validated gene predictions that do not overlap with the evidence-based catalogue.

The second step of filtering *de novo* gene predictions (by requiring the presence of adjacent exons in both species) turns out to greatly increase prediction specificity. Predicted genes that were removed by this criterion had a very low validation rate. In a sample of 101 predictions that failed to meet the criteria, the validation rate was 11% for genes with strong homology to human sequence and 3% for those without. The filtering process thus removed 24-fold more apparent false positives than true positives. Extrapolating from these results, testing the entire set of such predicted genes (that is, those that fail the test of having adjacent homologous exons in the two species) would be expected to yield only about 231 additional validated predictions.

Overall, we expect that about 1,000 (788+231) of the new gene predictions would be validated by RT–PCR. This probably corresponds to a smaller number of actual new genes, because some of these may belong to the same transcription unit as an adjacent *de novo* or evidence-based prediction. Conversely, some true genes may fail to have been detected by RT–PCR owing to lack of sensitivity or tissue, or developmental stage selection[327].

An example of a new gene prediction, validated by RT–PCR, is a homologue of dystrophin (Fig. 16). Dystrophin is encoded by the *DMD* gene, which is mutated in individuals with Duchenne muscular dystrophy[164]. A gene prediction was found on mouse chromosome 1 and human chromosome 2, showing 38% amino acid identity over 36% of the dystrophin protein (the carboxy terminal portion, which interacts with the transmembrane protein β-dystroglycan). Other new gene predictions include homologues of aquaporin. These gene predictions were missed by the evidence-based methods because they were below various thresholds. These and other examples are described in a companion paper[327].

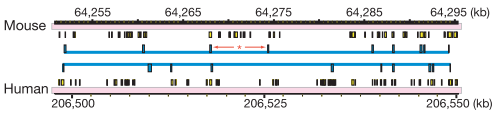


Figure 16 Structure of a new homologue of dystrophin as predicted on mouse chromosome 1 and human chromosome 2. Mouse and human gene structures are shown in blue on the chromosomes (pink). The mouse intron marked with an asterisk was verified by RT–PCR from primers complementary to the flanking exons followed by direct product sequencing[327]. Regions of high-scoring alignment to the entire other genome (computed before gene predictions and identification of predicted orthologues) are shown in yellow. Note the weak correspondence between predicted exons and blocks of high-scoring whole-genome alignment. Nonetheless, the predicted proteins considered in isolation show good alignment across several splice sites.

The overall results of the *de novo* gene prediction are encouraging in two respects. First, the results show that *de novo* gene prediction on the basis of two genome sequences can identify (at least partly) most predicted genes in the current mammalian gene catalogues with remarkably high specificity and without any information about cDNAs, ESTs or protein homologies from other organisms. It can also identify some additional genes not detected in the evidence-based analysis. Second, the results suggest that methods that avoid some of the inherent biases of evidence-based gene prediction do not identify more than a few thousand additional predicted exons or genes. These results are thus consistent with an estimate in the vicinity of 30,000 genes, subject to the uncertainties noted above.

RNA genes

The genome also encodes many RNAs that do not encode proteins, including abundant RNAs involved in mRNA processing and translation (such as ribosomal RNAs and tRNAs), and more recently discovered RNAs involved in the regulation of gene expression and other functions (such as micro RNAs)[165,166]. There are probably many new RNAs not yet discovered, but their computational identification has been difficult because they contain few hallmarks. Genomic comparisons have the potential to significantly increase the power of such predictions by using conservation to reveal relatively weak signals, such as those arising from RNA secondary structure[167]. We illustrate this by showing how comparative genomics can improve the recognition of even an extremely well understood gene family, the tRNA genes.

In our initial analysis of the human genome[1], the program tRNAscan-SE[168] predicted 518 tRNA genes and 118 pseudogenes. A small number (about 25 of the total) were filtered out by the RepeatMasker program as being fossils of the MIR transposon, a long-dead SINE element that was derived from a tRNA[169,170].

The analysis of the mouse genome is much more challenging because the mouse contains an active SINE (B2) that is derived from a tRNA and thus vastly complicates the task of identifying true tRNA genes. The tRNAscan-SE program predicted 2,764 tRNA genes and 22,314 pseudogenes in mouse, but the RepeatMasker program classified 2,266 of the 'genes' and 22,136 of the 'pseudogenes' as SINEs. After eliminating these, the remaining set contained 498 putative tRNA genes. Close analysis of this set suggested that it was still contaminated with a substantial number of pseudogenes. Specifically, 19 of the putative tRNA genes violated the wobble rules that specify that only 45 distinct anticodons are expected to decode the 61 standard sense codons, plus a selenocysteine tRNA species complementary to the UGA stop codon[171]. In contrast, the initial analysis of the human genome identified only three putative tRNA genes that violated the wobble rules[172,173].

To improve discrimination of functional tRNA genes, we exploited comparative genomic analysis of mouse and human. True functional tRNA genes would be expected to be highly conserved. Indeed, the 498 putative mouse tRNA genes differ on average by less than 5% (four differences in about 75 bp) from their nearest human match, and nearly half are identical. In contrast, non-genic tRNA-related sequences (those labelled as pseudogenes by tRNAscan-SE or as SINEs by RepeatMasker) differ by an average of 38% and none is within 5% divergence. Notably, the 19 suspect predictions that violate the wobble rules show an average of 26%

Curriculum Vitae

Name: Genís Parra Farré

Date and place of birth: June 6th, 1975. Barcelona, Catalonia (Spain)

Address: Genome Bioinformatic Lab. GRIB (IMIM/UPF/CRG)
Passeig Maritim de la Barceloneta, 37-49
08003 Barcelona, Catalonia (Spain)

Phone: +34 93 224 08 85

e-mail: gparra@imim.es

Web page: http://www1.imim.es/~gparra

Education

- Thesis title: **"Computational identification of genes: *ab initio* and comparative approaches"**
  Ph.D. advisor: Roderic Guigó
  At the Genome Bioinformatic Lab, Univertat Pompeu Fabra (2004).

- B.S. in Biology, specialization in Biomedicine, University of Barcelona (1998).

Publications:

- International Chicken Genome Sequencing Consortium (including **G. Parra** and R. Guigó). **"Sequencing and comparative analysis of the chicken genome"**. *Nature*, in press.

- International Tetraodon Genome Sequencing Consortium (including **G. Parra** and R. Guigó). **"Duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype"**. *Nature*, 431:946-957 (2004).

- **G. Parra**, P. Agarwal, J.F. Abril, T. Wiehe, J.W. Fickett and R. Guigó. **"Comparative gene prediction in human and mouse"**. *Genome Research* 13(1):108-117 (2003).

- R. Guigó, E.T. Dermitzakis, P. Agarwal, C.P. Ponting, **G. Parra**, A. Reymond, J.F. Abril, E. Keibler, R. Lyle, C. Ucla, S.E. Antonarakis and M.R. Brent. **"Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes"**. *Proc. Nat. Acad. Sci.* 100(3):1140-1145 (2002).
- Mouse Genome Sequencing Consortium (including **G. Parra** and R. Guigó). **"Initial sequencing and comparative analysis of the mouse genome"**. *Nature* 420(6915):520-562 (2002).
- G. Glökner, L. Eichinger, K. Szafranski, J.A. Pachebat, A.T. Bankier, P.H. Dear, R. Lehmann, C. Baumgart, **G. Parra**, J.F. Abril, R. Guigó, K. Kumpf, B. Tunggal, the Dictyostelium Genome Sequencing Consortium, E. Cox, M.A. Quail, M. Platzer, A. Rosenthal and A.A. Noegel. **"Sequence and Analysis of Chromosome 2 of Dictyostelium discoideum"**. *Nature* 418(6893):79-85 (2002).
- **G. Parra**, E. Blanco, y R. Guigó. **"geneid in Drosophila"**. *Genome Research*, 10, 511-515 (2000).

Book chapters:

- E. Blanco, **G. Parra** and R. Guigó. Using geneid to Identify Genes. In A. Baxevanis, editor: Current Protocols in Bioinformatics. Unit 4.3. John Wiley & Sons Inc., New York (2002).

Posters:

- **G. Parra**, A. Reymond, N. Dabbouseh, S.E. Antonaraquis, T.M. Thomson and R. Guigó. "Tandem chimerism in the human genome" Genome Informatics. Cold Spring Harbor Laboratory (2004, Hinxton, UK).
- G. Glöckner, L. Eichinger, K. Szafranski, P. Dear, J. Pachebat, K. Kumpf, R. Lehmann, J.F. Abril, **G. Parra**, R. Guigó, B. Tunggal, E. Cox, M.A. Quail, M. Platzer, A. Rosenthal, A.A. Noegel and the Dictyostelium Genome Sequencing Consortium. "Sequence and analysis of chromosome 2 from the model organism *Dictiostelium Discoideum*" Genome Sequencing and Biology. Cold Spring Harbor Laboratory (2001, New York, USA).
- E. Blanco, **G. Parra**, J.F. Abril, M. Burset, S. Castellano, X. Fustero, y R. Guigó; " Gene predictions in the post-genomic era" ISMB (2001, Copenhagen. Denmark).
- J.F. Abril, E. Blanco, M. Burset, S. Castellano, X. Fustero, **G. Parra** and R. Guigó; "Genome Informatics Research Laboratory: Main Research Topics." *I Jornadas de Bioinformática* (2000, Cartagena, Spain).

Teaching experience:

- Master in Bioinformatics and Computational biology (10h). Universidad Complutense de Madrid (2004). Madrid (Spain).

- Graduate Programme in Bioinformatics. Gene Prediction and Identification (40h). Faculty of Sciences of the University of Lisbon, (FCUL) and the Gulbenkian Institute of Science of the Calouste Gulbenkian Foundation (IGC) (2002-2003). Oeiras (Portugal).
- A Bioinfomatic course (60h). Universitat Pompeu Fabra (2002-2003). Barcelona, Catalonia (Spain).
- Computational gene identification (10h). Gulbenkian Institute of Science of the Calouste Gulbenkian Foundation (IGC) (2002). Oeiras (Portugal).
- Bioinformatics for Comparative and Functional Genomics (2h). EMBL course. Universitat Pompeu Fabra (2001). Barcelona, Catalonia (Spain).

Visiting research positions:

- February - March 2002. Under the supervision of Pankaj Agarwal. GlaxoSmithKline. King of Prussia, Pennsylvania (USA).
- December 2001. Under the supervision of Gernot Glökner. Department of Genome Analysis. Institute of Molecular Biology. Jena (Germany).

Related experience:

- September 1998 - July 1999. Computer technical support in the Faculty of Biology, Universitat de Barcelona.

Bibliography

Alexandersson, M., Cawley, S., and Pachter, L. (2003). SLAM: cross-species gene finding and alignment with a generalized pair hidden markov model. *Genome Research*, 13:496–502.

Arthur, J. and Wilkins, M. (2004). Using proteomics to mine genome sequences. *Journal of proteome research*, 3:393–402.

Ashburner, M., Misra, J., Roote, J., Lewis, S., Blazej, R., andC. Doyle, T. D., George, R. G. R., n. Harris, g. Hartzell, d. Harvey, l. Hong, k.Houston, Hoskins, R., Johnson, G., Martin, C., Moshrefi, A., Palazzolo, M., Reese, M., Spradling, A., Tsang, G., Wan, K., Whitelaw, K., Celniker, S., and et al (1999). An exploration of the sequence of a 2.9 Mb region of the genome of drosophila melanogaster. the adh region. *Genetics*, 153:179–219.

Bafna, V. and Huson, D. H. (2000). The conserved exon method. *Proceedings of the eigth intenational conference on Intelligent Systems in Molecular Biology (ISMB)*, pages 3–12.

Bajic, V. and Seah, S. (2003). Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcroptional units. *Genome Research*, 13:1923–1929.

Batzoglou, S., Patcher, L., Mesirov, J. P., Berger, B., and Lander, E. S. (2000). Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Research*, 10(7):950–958.

Beltran, S., Blanco, E., Serras, F., Perez-Villamil, B., Guigó, R., Artavanis-Tsakonas, S., and Corominas, M. (2003). Transcriptional network controlled by the trithorax-group gene ash2 in drosphila melanogaster. *Proceedings National Academy Sciences USA*, 100:3293–3298.

Birney, E., Andrews, T., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K., Cameron, G., Durbin, R., Cox, A., Hubbard, T., and Clamp, M. (2004a). An overview of ensembl. *Genome Research*, 14:925–928.

Birney, E., Clamp, M., and Durbin, R. (2004b). Genewise and genomewise. *Genome Reseach*, 14:988–995.

Blanco, E. (2000). Diseño de aplicaciones paralelas para la predicción de elementos génicos. Master's thesis, Universitat Politècnica de Catalunya.

Blayo, P., Rouzé, P., and Sagot, M.-F. (2002). Orphan gene finding - an exon assembly approach. *Theoretical Computer Science. in press*.

Borodovsky, M. and McIninch, J. (1993). GeneMark: Parallel gene recognition for both DNA strands. *Computer and Chemistry*, 17:123–134.

Burge, C. B. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94.

Burge, C. B. and Karlin, S. (1998). Finding the genes in genomic DNA. *Current Opinion in Structural Biology*, 8:346–354.

Burge, C. B., Tuschl, T., and Sharp, P. S. (1999). Splicing precursors to mrnas by the spliceosomes. In Gesteland, R. F., Cech, T. R., and Atkins, J. F., editors, *The RNA world*, pages 525–560, Cold Spring Harbor, New York. Cold Spring Harbor Laboratory Press.

Burset, M. and Guigó, R. (1996). Evaluation of gene structure prediction programs. *Genomics*, 34:353–357.

Cartegni, L., Chew, S. L., and Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature reviews genetics*, 3:285–298.

Castellano, S. (2004). Towards the characterization of the eukaryotic selenoproteome. *GBL diseestation series*, 2004-01.

Castellano, S., Morozova, N., Morey, M., Berry, M., Serras, F., Corominas, M., and Guigó, R. (2001). "*in silico*" identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Reports*, 2:697–702.

Castellano, S., Novoselov, S. V., Kryukov, G. V., Lescure, A., Blanco, E., Krol, A., Gladyshev, V. N., and Guigó, R. (2004). Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution. *EMBO Rep.*, 5:71–77.

Castelo, R. and Guigó, R. (2004). Splice site identification by idlBNs. *Bioinformatics*, 2:697–702.

Coelho, P., Bryan, A., Kumar, A., and G.S. Shadel, M. S. (2002). A novel mitochondrial protein, tar1p, is encoded on the antisense strand of the nuclear 25s rdna. *Genes and developement*, 16:2755–60.

Curwen, V., Eyras, E., Andrews, T., Clarke, L., Mongin, E., Searle, S., and Clamp, M. (2004). The ensembl automatic gene annotation system. *Genome Research*, 14:942–950.

Das, M., Burge, C. B., Park, E., Colinas, J., and Pelletier, J. (2001). Assesment of the Total Number of Human Transcription Units. *Genomics*, 77(1–2):71–78.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., and Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acid Research*, 27(11):2369–2376.

Dewey, C., Wu, J., Cawley, S., Alexandersson, M., Gibbs, R., and Pachter, L. (2004). Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Research*, 14:661–664.

Dong, S. and Searls, D. B. (1994). Gene structure prediction by linguistic methods. *Genomics*, 23:540–551.

Down, T. and Hubbard, T. (2002). Computational detection and location of transcription start sites in mammalian genomic dna. *Genome Research*, 12:458–461.

Dunlop, J., Corominas, M., and Serras, F. (2000). The novel gene glaikit, is expressed during neurogenesis in the drosophila melanogaster embryo. *Mech Dev.*, 96:133–136.

Durbin, R., Eddy, S., Crogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids*. Cambridge University Press.

Eyras, E., Caccano, M., Curwen, V., and Clamp, M. (2004). ESTGenes: alternative splicing from ESTs in ENSEMBL. *Genome Research*, 14:976–987.

Fazzari, M. and Greally, J. (2004). Epigenomics: beyond CpG islands. *Nature reviews*, 5:446–454.

Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, 10:5303–5318.

Fickett, J. W. and Tung, C.-S. (1992). Assessment of protein coding measures. *Nucleic Acids Research*, 20:6441–6450.

Fields, C. A. and Soderlund, C. A. (1990). gm: a practical tool for automating DNA sequence analysis. *Computer Applications to the Biosciences*, 6:263–270.

Galperin, M. Y. and Koonin, E. V. (2003). In *Frontiers in Computational Genomic*. Caister Academic Press.

Gelfand, M. S. (1990). Computer prediction of exon-intron structure of mammalian pre-mRNAs. *Nucleic acid research*, 18:5865–5869.

Gelfand, M. S. (1995). Prediction of function in DNA sequence analysis. *Journal of Computational Biology*, 1:87–115.

Gelfand, M. S. and Roytberg, M. A. (1993). Prediction of the exon-intron structure by a dynamic programming approach. *BioSystems*, 30:173–182.

Glokner, G., Eichinger, L., Szafranski, K., Pachebat, J., Bankier, A., Dear, P., Lehmann, R., Baumgart, C., Parra, G., Abril, J., Guigo, R., Kumpf, K., Tunggal, B., the Dictyostelium Genome Sequencing Consortium, Cox, E., Quail, M., Platzer, M., Rosenthal, A., and Noegel, A. (2002). Sequence and analysis of chromosome 2 of dictyostelium discoideum. *Nature*, 418:79–85.

Graveley, B. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17:100–107.

Guigó, R. (1998). Assembling genes from predicted exons in linear time with dynamic programming. *Journal of Computational Biology*, 5:681–702.

Guigó, R. (1999). DNA composition, codon usage and exon prediction. In Bishop, M., editor, *Genetic Databases*, pages 53–80. Academic Press, San Diego, California.

Guigó, R., Agarwal, P., Abril, J. F., Burset, M., and Fickett, J. W. (2000). Gene prediction accuracy in large DNA sequences. *Genome Research*, 10:1631–1642.

Guigó, R., Knudsen, S., Drake, N., and Smith, T. F. (1992). Prediction of gene structure. *Journal of Molecular Biology*, 226:141–157.

Henderson, J., Salzberg, S., and Fassman, K. H. (1997). Finding genes in DNA with a hidden markov model. *Journal of Computational Biology*, 4:127–141.

Howe, K., Chothia, T., and Durbin, R. (2002). Gaze: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Research*, 12:1418–1427.

Hutchinson, G. B. and Hayden, M. R. (1992). The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Research,*, 20:3453–3462.

Johnson, J., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P., Santos, C. A. R., Schadt, E., Stoughton, R., and Shoemaker, D. (2003). Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *science*, 302:2141–2144.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of American Statistical Association*, 90:773–795.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5:59.

Korf, I., Flicek, P., Duan, D., and Brent, M. R. (2001). Integrating Genomic Homology into Gene Structure Prediction. *Bioinformatics*, 17(1):S140–8.

Kozak, M. (1987). An analysis of 5'-noncodingsequences from 699 vertebrate messenger rnas. *Nucleic Acids Research*, 15:8125–8148.

Kozak, M. (1999). Initiation of translation in prokariotes and eukaryotes. *Gene*, 234:187–208.

Krogh, A. (1997). Two methods for improving performance of an HMM and their application for gene finding. *ISMB*, 5:179–186.

Kryukov, G. V., Castellano, S., Novoselov, S. V., Lobanov, A. V., Zehtab, O., Guigó, R., and Gladyshev, V. N. (2003). Characterization of mammalian selenoproteomes. *Science*, 300:1439–1443.

Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H. (1996). A generalized hidden markov model for the recognition of human genes in DNA. In States, D. J., Agarwal, P., Gaasterland, T., Hunter, L., and Smith, R., editors, *Intelligent Systems for Molecular Biology*, pages 134–142, Menlo Park, California. AAAI press.

Levine, M. and Tjian, R. (2003). Transcrition regulation and animal diversity. *Nature*, 424:147–150.

Lewis, B. (1997). Genes VI. Great Clarendon Street, Oxford. Oxford University Press.

Lowe, T. and Eddy, S. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acid research*, 25:955–964.

Lukashin, A. V. and Borodovski, M. (1998). genemark.hmm: new solutions for gene finding. *Nucleic acid research*, 26:1107–1115.

Meyer, I. and Durbin, R. (2002). Comparative *ab initio* prediction of gene structure using pair hmms. *Bioinformatics*. *in press*.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, 20:13–19.

Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Computer Applications in the Biosciences*, 13:477–478.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562.

Notredame, C., Higgins, D., and Heringa, J. (2000). T-coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology*, 302:205–217.

Parra, G., Agarwal, P., Abril, J. F., Wiehe, T., Fickett, J. W., and Guigó, R. (2003). Comparative gene prediction in human and mouse. *Genome Research*.

Parra, G., Blanco, E., and Guigó, R. (2000). Geneid in *Drosophila*. *Genome Research*, 10:511–515.

Pedersen, C. and Scharl, T. (2002). Comparative methods for gene structure prediction in homologou s sequences. In Guigó, R. and Gusfield, D., editors, *Algorithms in Bioinformatics*.

Pedersen, J. and Hein, J. (2003). Gene finding with a hidden markov model of genome structure and evolution. *Bioinformatics*, 22:219–227.

Penn, S., Rank, D., Hanzel, D., and Barker, D. (2000). Mining the human genome using microarrays of open reading frames. *Nature genetics*, 26:315–318.

Poulin, F., Brueschke, A., and Sonenberg, N. (2003). Gene fusion and overlapping reading frames in the mammalian genes for 4e-bp3 and mask. *J. Biol. Chem.*, 278:52290–52297.

Rabinier, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286.

Reese, M. G., Eeckman, F. H., Kulpand, D., and Haussler, D. (1997). Improved splice site detection in Genie. *J Comput Biol*, 4(3):311–323.

Reese, M. G., Hartzell, G., Harris, N. L., Ohler, U., Abril, J. F., and Lewis, S. E. (2000). Genome annotation assessment in *Drosophila melanogaster*. *Genome Research*, 10:483–501.

Rogic, S., Mackworth, A. K., and Ouellette, F. (2001). Evaluation of gene-finding programs on mammalian sequences. *Genome Research*, 11:817–832.

Roy, S. W. (2003). Recent evidence for the exon theory of genes. *Genetica*, 118:251–266.

Salamov, A. and Solovyev, V. (2000). Ab initio gene finding in drosophila genome DNA. *Genome Research*, 10:516–522.

Senapathy, P., Shapiro, M., and Harris, N. (1990). Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods in Enzymology*, 183:252–278.

Sharp, P. A. and Burge, C. B. (1997). Classification of introns: U2-type or U12-type. *Cell*, 91:875–879.

Shepherd, J. C. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proceedings National Academy Sciences USA.*, 78:1596–1600.

Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garret-Engele, P., McDonagh, P. D., Loerch, P. M., Leonardson, A., Lum, P. Y., Cavet, G., Wu, L. F., Altschuler, S. J., Edwards, S., King, J., Tsang, J. S., Schimmack, G., Schelter, J. M., Koch, J., Ziman, M., Marton, M. J., Li, B., Cundiff, P., Ward, T., Castle, J., Krolewski, M., Meyer, M. R., Mao, M., Burchard, J., Kidd, M. J., Dai, H., Phillips, J. W., Linsley, P. S., Stoughton, R., Scherer, S., and Bogusky, M. S. (2001). Experimental annotation of the human genome using microarray tecnology. *Nature*, 409(6822):922–927.

Siepel, A. and Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology Evolution*, 21:468–488.

Smit, A. and Green, P. (1999). RepeatMasker at `http://ftp.genome.washington.edu/rm/ repeatmasker.html`. *unpublished*.

Snyder, E. E. and Stormo, G. D. (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Research*, 21:607–613.

Snyder, M. and Gerstein, M. (2003). Defining genes in the genomic era. *Science*, 300:258–260.

Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research*, 22:5156–5163.

Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. (1995). Identification of human gene structure using linear discriminant functions and dynamic programming. *Proc Int Conf Intell Syst Mol Biol.*, 3:367–75.

Staden, R. and McLachlan, A. D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research*, 10:141–156.

Tenney, A. E., Brown, R. H., Vaske, C., Lodge, J. K., Doering, T. L., and Brent, M. R. (2004). Gene prediction and verification in a compact genome with numerous small introns. *Genome Reseach. in press*.

Tetraodon Genome Sequencing Consortium (2004). Analysis of the tetraodon nigroviridis genome reveals the protolkaryotpe of bony vertebrates and its duplication in teleost fish. *Nature. in press*.

Thomas, A. and Skolnick, M. H. (1994). A probabilistic model for detecting coding regions in DNA sequences. *IMA Journal of Mathematical Applications to Medicine and Biology*, 11:149–160.

Thomson, T., Lozano, J. J., Carrió, R., Serras, F., Loukili, N., Valeri, M., Cormand, B., del Río, M. P., Abril, J. F., Burset, M., Sancho, E., Merino, J., Macaya, A., Corominas, M., and Guigó, R. (2000). Fusion of the human gene for the polyubiquitination co-effector UEV-1 with *Kua*, a newly identified gene. *Genome Research*, 10:1743–1756.

Tycowski, K., Shu, M., and Steitz, J. (1996). A mammalian gene with introns instead of exons generating stable rna products. *Nature*, 379:464–466.

Uberbacher, E. C. and Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proceedings National Academy Sciences USA*, 88:11261–11265.

Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T., and Guigó, R. (2001). SGP-1: Prediction and validation of homologous genes based on sequence alignments. *Genome Research*, 11:1574–1583.

Xu, Y., Mural, R. J., and Uberbacher, E. C. (1994). Constructing gene models from accurately predicted exons: An application of dynamic programming. *Computer Applications in the Biosciences*, 11:117–124.

Zhang, M. Q. (2002). Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, 3:698–709.

Zhang, M. Q. and Marr, T. (1993). A weight array method for splicing signal analysis. *CABIOS*, 9:499–509.

Zorio, D. and Bentley, D. (2004). The link between mRNA processing and transcription: communication works both ways. *Experimental Cell Research*, 296:91–97.

Notes

Titles in the GBL dissertation series

2002-01 Moisés Burset.
Estudi computacional de l'especificació dels llocs d'splicing.
Departament de Genètica, Universitat de Barcelona.

2004-01 Sergi Castellano.
Towards the characterization of the eukaryotic selenoproteome: a computational approach.
Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra.

2004-02 Genís Parra.
Computational identification of genes: ab initio and comparative approaches.
Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra.