COMPARATIVE GENOMICS: GENOME-WIDE ANALYSIS IN METAZOAN EUKARYOTES

Abel Ureta-Vidal, Laurence Ettwiller and Ewan Birney

The increasing number of complete and nearly complete metazoan genome sequences provides a significant amount of material for large-scale comparative genomic analysis. Finding new effective methods to analyse such enormous datasets has been the object of intense research. Three main areas in comparative genomics have recently shown important developments: whole-genome alignment, gene prediction and regulatory-region prediction. Each of these areas improves the methods of deciphering long genomic sequences and uncovering what lies hidden in them.

RATE MATRIX
Denotes the probability of mutation from one amino acid to another (or from one nucleotide to another) for a given period of evolution. The most well known rate matrices are BLOSUM and PAM.

EnsEMBL Project, Room A2–06, EMBL–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Correspondence to E.B. e-mail: birney@ebi.ac.uk
doi:10.1038/nrg1043

Over the past 4 years, the complete, or nearly complete, genome sequences from some of the most important metazoan organisms — including *Caenorhabditis elegans*[1], *Drosophila melanogaster*[2], *Homo sapiens*[3,4], *Fugu rubripes*[5], *Anopheles gambiae*[6], *Mus musculus*[7], *Caenorhabditis briggsae*, *Rattus norvegicus* and *Ciona intestinalis*[8] — have been determined and made publicly available (FIG. 1). Several other genome sequences will be released over the next 2 years, including *Pan troglodytes*, *Danio rerio*, *Tetraodon nigroviridis*, *Oryzias latipes*, *Gallus gallus* and *Ciona savignyi*. Comparing genomes provides a tremendous insight into the biology of the organisms that are examined and into the process of evolution. Recent papers on the genomes of *A. gambiae* (compared to *D. melanogaster*)[9,10] and *M. musculus* (compared to *H. sapiens*)[7,11,12] show the impressive power of this approach. This review examines some of the issues in metazoan comparative genomics, and its impact on two areas of functional prediction: the prediction of gene structure and of *cis*-acting regulatory elements. We do not describe other aspects of comparative genomics, such as phylogeny[13] and molecular evolution[14], although we include some of these in our discussion; neither do we describe comparative genomics in plants[15,16], archaea and eubacteria[17–21] — all of which are vibrant and fascinating fields in themselves.

Comparative genomics

Genome sequence comparison is a good example of the application of modern evolutionary theory, as advocated by Kimura[22] and others[23,24]. The assumption that underlies all comparative genomics is that the two genomes under examination had a common ancestor and, therefore, that every base pair in each organism can be explained as the combination of this original ancestral genome and the action of evolution. Evolution can be broadly thought of as the combination of two processes: mutational forces that generate random mutations in the genome sequence, and selection pressures that eliminate random mutations (negative selection), have no effect on mutations (neutral selection) or increase the frequency of mutant alleles in the population as a result of a gain in fitness (positive selection). The combined action of mutation and selection is represented generally by a RATE MATRIX of base-pair changes between the two observed genomes. This rate matrix is then used in several algorithmical approaches that attempt to reconstruct the 'most likely' explanation of the action of evolution. The most likely explanation generally provides an alignment of base pairs between the two genomes, with each base-pair alignment indicating a single original base pair in the ancestral genome. However, our understanding of these evolutionary processes is woefully inadequate to explain the observed data. Therefore, most work

© 2003 Nature Publishing Group

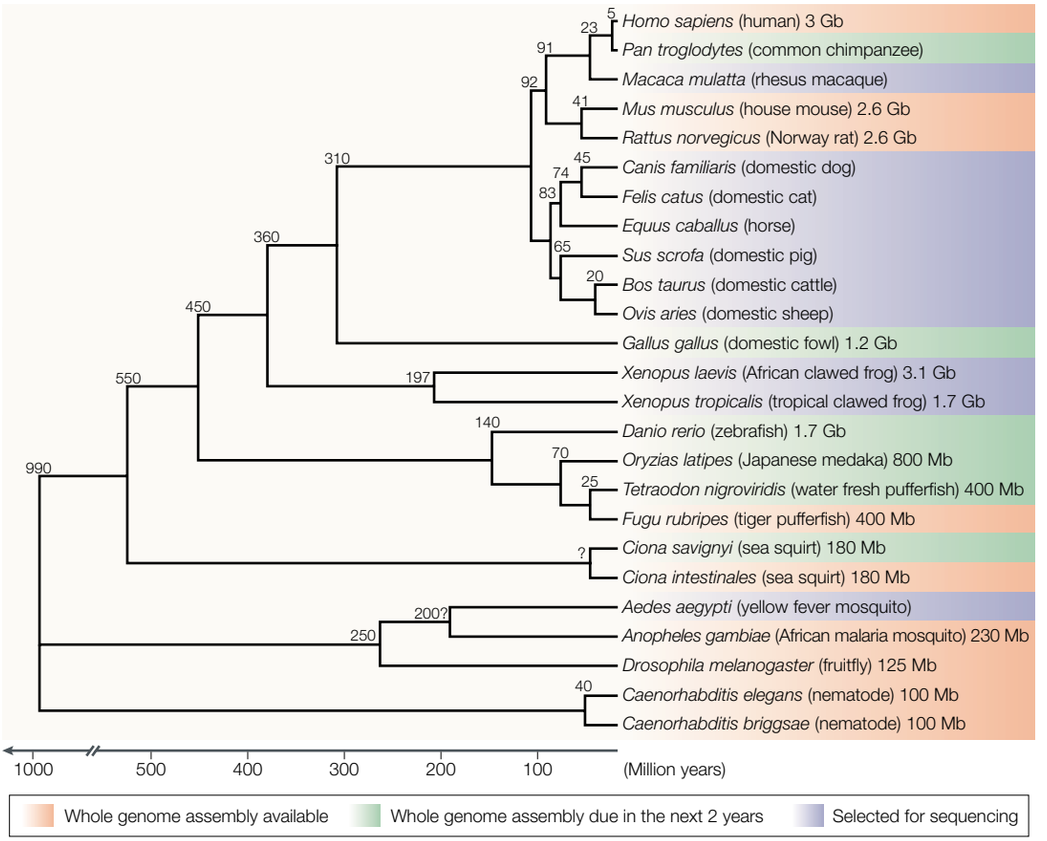


Figure 1 | **Evolutionary relationship between metazoans that are sequenced or due for sequencing.** The simplified phylogenetic relationships between the metazoans for which the complete, or nearly complete, genome sequences are available or will be available soon. Evolutionary distances (in million years) and genome sizes are based on REFS 8,10,13,98–100.

has focused on the practicalities of generating and analysing the data to gain insights into conserved FUNCTIONAL SEQUENCES, which, in general, means understanding the action of negative selection.

As well as such theoretical concerns, there are substantial practical problems to deal with. The size of the genomes, which range from 100 Mb to 3 Gb, means that the basic algorithmical approach (comparing every base in one genome to every base in another using a rigorous mathematical framework) is impractical. Indeed, just storing the genomic information and the results of the comparisons is a challenge. Genome sequences also differ in their state of readiness, from 'essentially finished' genomes (such as those of *C. elegans*, *D. melanogaster* and *H. sapiens*) to draft genomes (such as those of *A. gambiae* and *M. musculus*) and unassembled genomes that will be compiled at a later date (such as *D. rerio* and *Tetraodon*). Because of these practical problems, most biologists will access the precomputed results of large genome comparisons, instead of undertaking their own analyses.

In the past 5 years, an increasing number of programs have been produced that aim to align large genomic sequences from different organisms. Each of these attempts, in different ways, to address the problems of speed and accuracy for whole-genome alignments. Because this field is rapidly evolving and productive[25], the explosion in the availability of tools might confuse end-user biologists and make it difficult for them to decide which tool is most suitable for their research. Indeed, to interpret these resources, biologists must be aware of the differences between them and between their underlying assumptions.

For individual genomic regions, 'adventurous' biologists might do their own comparisons and interpretation. This is generally advisable, because the large-scale genome comparisons use algorithms that are necessarily less sensitive than those that search for the optimal alignment[26,27] and, perhaps more importantly, because it is only by engaging with the data and the computational methods that it is possible to have a feel for them.

FUNCTIONAL SEQUENCE
A genomic sequence that provides a function that is under selection and tends to be conserved between species. For example, a protein-coding region or transcription-factor binding site

Whole-genome alignments

Precomputed alignments. So far, several groups have made large cross-species comparisons (FIG. 2), and released them for downloading or for visualization in a genome browser. At the time of writing, updated human–mouse comparisons could be obtained from three groups: the University of California Santa Cruz (UCSC)/PennState (translated BLAT or BLASTZ), the

© 2003 Nature Publishing Group

Berkeley Genome Pipeline (using BLAT/AVID) and Ensembl (using Phusion/Blastn). Comparisons of the rat assembly with the human and mouse sequences have also recently been released by these three groups. Human genome comparisons to either mouse or *Fugu* can be viewed using the ECR browser, which is based on VISTA[28], at Berkeley. Finally, the WABA (see below) and Phusion/Blastn alignments of *C. elegans* and *C. briggsae* are accessible through the WormBase[29] and Ensembl[30] genome browsers, respectively.

These data can be visualized using various interactive graphical interfaces, including the VISTA, UCSC[31] and Ensembl genome browsers (BOX 1). To put this comparative data into perspective with other information, such as the location of a gene or a repeat, biologists need an easy way of accessing the data and setting simple tasks, such as "find all the highly conserved regions between human and mouse that are located 1,000 bp 5′ of the transcription start site" or "find all alignments in a particular set of ancient repeats". The GALA[32] and EnsMart[30] data-mining systems are good examples of such functionality.

Which genomes to align? The following basic assumptions are made when two genomes are aligned: that they are derived from a common ancestor, that the time since their divergence has enabled a significant accumulation of mutations in the two genomes and that selection has occurred. However, sufficient similarity between the genomes must remain to enable the easy identification of homologous regions. So, DNA alignments between human and mouse, which diverged ~90 million years ago (Mya) (FIG. 1), have been useful in many cases[7,11,12,33–35] and have led to, for example, the finding of new genes such as *APOAV*[33] or of gene regulatory regions[34]. Similarly, alignments between human and *Fugu*/*Tetraodon* (which diverged 450 Mya), although less easy, owing to the long divergence time, are still feasible and have produced the same kind of interesting information[36,37].

Comparisons between *C. elegans* and *C. briggsae* (which diverged 40 Mya)[38] have also been carried out. *Drosophila* and *Anopheles* have accumulated a large number of mutations and rearrangements since their divergence from a common ancestor 250 Mya. Therefore, a whole-genome comparison at the nucleotide level has been carried out using the classical TBLASTX, not as an attempt to properly align the two genomes, but, as part of the Celera Annotation Pipeline, to predict genes[6]. Unfortunately, the results of such comparisons are not available. Instead, comparisons have been made at the protein level[10], but the value of this approach relies on having comprehensive and validated sets of protein-coding regions — a requirement that might be satisfied for *Drosophila* but not for *Anopheles*. By comparing genomes only at the protein level, information is lost that could help to identify new coding or non-coding genes, as well as regulatory regions. In the first instance, aligning closely related genomes that belong to the same branch of the evolutionary tree — for example, vertebrates, arthropods and nematodes — is the best way to start.

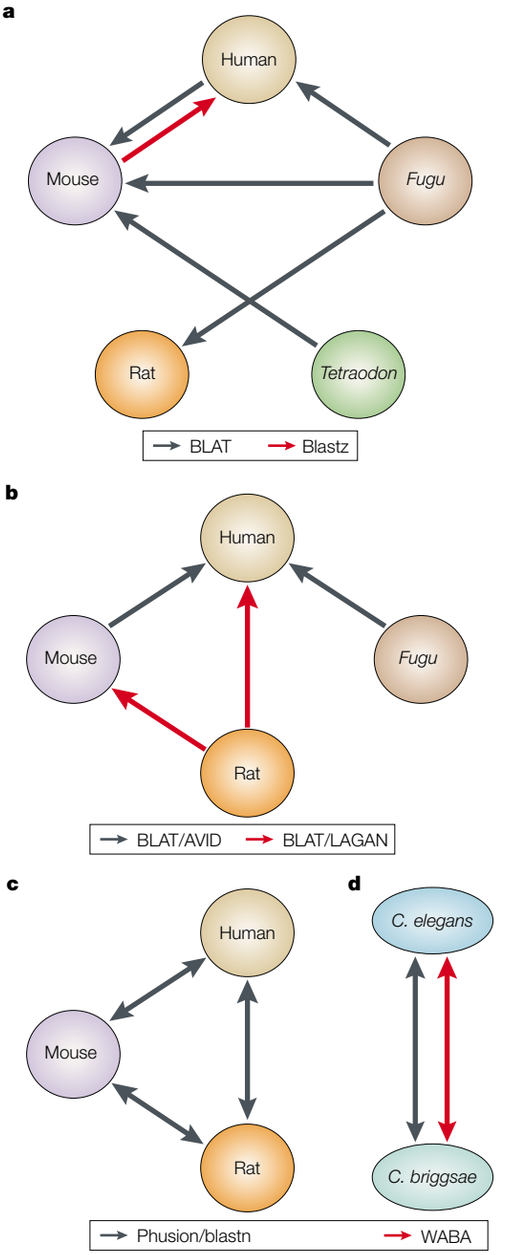


Figure 2 | **Whole-genome alignments available online.** Whole-genome precomputed alignments are available from the web sites of several different groups. Each group uses a different alignment method. The figure gives an overview of what is provided by each group and the methods used. The direction of the arrows indicate the species that are used as the reference for the display. For example, in the rat–mouse pair in panel **b**, the arrow pointing to mouse indicates that rat alignments can be seen on the mouse genome (but not the opposite), and in the human–mouse pair in panel **c**, the arrows pointing to both species indicate that human alignments can been seen on the mouse genome, and vice versa. Alignments can be found at the University of California, Santa Cruz, UCSC Genome Browser and the Penn State Bioinformatics Group (**a**), the Berkeley Genome Pipeline and the ECR Browser (**b**), the Ensembl Genome Browser (**c** and **d**) and WormBase (**d**). The figure represents a snapshot of the alignments that are available at the time of writing.

© 2003 Nature Publishing Group

SEEDS
A short exact, or nearly exact, matching string od characters aligning between two sequences.

Alignment strategies. The naive approach for calculating an alignment of two sequences is to create a matrix using a recursive calculation called dynamic programming (DP) (BOX 2). For short DNA sequences and protein sequences this is feasible and relies on two well-known algorithms: Needleman and Wunsch (NW)[27] and Smith-Waterman (SW)[26], both of which have two main limitations. First, they do not properly handle alignments of rearranged sequences (such as inversions, duplications and translocations). Second, and most important, if the two sequences are >10,000 bp, calculating alignments by these methods becomes extremely expensive both in terms of time and memory usage, and is not feasible. To gain speed, all fast alignment programs rely on some sort of heuristic (or seeding) strategy (BOX 3) to reduce the computational complexity. A seeding strategy is based on the idea that correct alignments will have significant stretches of ungapped matches; by first finding a set of significant ungapped matches, the algorithm can avoid large amounts of computation. The loss of sensitivity of such heuristics is tolerated, as it gives a tremendous gain in speed and a reduction in memory usage.

Generally, all algorithms go through three main steps, although how these steps are executed differs depending on the algorithm. In step one, SEEDS are found using the seeding strategy that is particular to a given algorithm. In step two, the seeds can be used as nucleation points for extension or as anchor points to partition the alignment problem. In step three, a DP process is used to produce gapped alignments (FIG. 2). Here, we concentrate on four whole-genome alignment approaches: BLASTZ, BLAT/AVID, BLAT/LAGAN and WABA. Other approaches exist, including MUMmer[39,40], PatternHunter[41], GLASS[42], MegaBLAST (BOX 1) and Phusion/Blastn[43,44], all of which work in a similar three-step way.

Local aligners, such as BLASTZ and BLAT, are designed to be highly sensitive but less specific, and according to the scoring scheme and seeding strategy used, they will find all possible sequence similarities, not just the contiguous ones. BLASTZ[45] uses several simple methods to increase its specificity and sensitivity, and applies the basic three-step alignment strategy used by Gapped BLAST[44]: find the seeds, extend the seeds without gaps and extend further by allowing gaps using DP (FIG. 3).

BLASTZ can handle large sequences — for example, human/mouse sequences of 10–30 Mb. In BLASTZ, the basic search for seeds looks for exact short matches (usually 11 bases in the classical BLAST). A new approach, initially implemented in PatternHunter[41] — which is another genome aligner — involves looking for nearly exact matches, using a two weighted-spaced-seeds strategy (BOX 3). This approach identifies significantly more seeds than the classical consecutive model, and, therefore, expands the range of the alignment to more regions. Formerly, BLASTZ used the consecutive model for seeding (BOX 3), but the two weighted-spaced seeds (12of19) have now been implemented to increase its sensitivity. It also tolerates a transition (A–G, G–A, C–T or T–C) in 1 of the 12 positions. In the seeding step, when many regions of the query sequence are aligned with the same region of the database sequence, these repeated regions are masked and are not used in the following steps, which increases the speed and specificity. After the gap-free extension step, low-complexity matches (such as CACACACA) have their score specifically down-weighted[46] to reduce the possibility of a gapped alignment in the third step, further increasing the specificity and speed. The three steps are repeated

Box 1 | **Internet resources**

Genome aligners

AVID	http://www-gsd.lbl.gov/vista/details_avid.htm
BLASTZ	http://bio.cse.psu.edu
BLAT	http://genome.ucsc.edu
Exonerate	http://www.ebi.ac.uk/~guy/exonerate
GLASS	http://crossspecies.lcs.mit.edu
LAGAN/MLAGAN	http://lagan.stanford.edu
MegaBLAST	http://www.ncbi.nih.gov/blast/tracemb.html
MUMmer	http://www.tigr.org/software/mummer
PatternHunter	http://www.bioinformaticssolutions.com/
	products/ph.php
WABA	http://www.cse.ucsc.edu/~kent/xenoAli/index.html

Prediction of exons or coding regions

DIALIGN2	http://bibiserv.techfak.uni-bielefeld.de/dialign
ExoFish	http://www.genoscope.cns.fr/proxy/cgi-bin/exofish.cgi
OrthoSeq	http://www.phylofoot.org/cgi-bin/orthoseq.cgi
ROSETTA/GLASS	http://crossspecies.lcs.mit.edu

Prediction of exons and gene structure

DoubleScan	http://www.sanger.ac.uk/Software/analysis/doublescan
SLAM	http://baboon.math.berkeley.edu/~syntenic/slam.html
SPG-1	http://soft.ice.mpg.de/sgp-1
TwinScan	http://genes.cs.wustl.edu

Finding regulatory regions by phylogenetic footprinting

Consite	http://forkhead.cgb.ki.se/cgi-bin/consite
Footprinter	http://abstract.cs.washington.edu/~blanchem/
	FootPrinterWeb/FootprinterInput.pl
rVISTA	http://nemo.lbl.gov/rvista/index.html
Toucan	http://www.esat.kuleuven.ac.be/~saerts/
	software/toucan.php
Trafac	http://trafac.chmcc.org/trafac/index.jsp
Toucan	http://www.esat.kuleuven.ac.be/~saerts/
	software/toucan.php

Whole-genome alignments in genome browser

ECR browser	http://nemo.lbl.gov/ecrBrowser
Ensembl	http://www.ensembl.org
UCSC	http://genome.ucsc.edu
Pipmaker (pdf output only)	http://bio.cse.psu.edu/pipmaker
VISTA-based	http://pipeline.lbl.gov
WormBase	http://www.wormbase.org

Visualization tools

Alfresco	http://www.sanger.ac.uk/Software/Alfresco
Aplot	http://www1.imim.es/~jabril/GFFTOOLS/APLOT.html
Apollo	http://www.ensembl.org/apollo
GenomePixelizer	http://www.atgc.org/GenomePixelizer/
	GenomePixelizer_Welcome.html
SynPlot	http://www.sanger.ac.uk/Users/jgrg/SynPlot

© 2003 Nature Publishing Group

once for regions that lie between matches and that share the same order and orientation if they are separated by <50 kb (or 10 kb, thereby allowing a lower threshold on matches) using a more sensitive seed (for example, a 7-mer). A further constraint is subsequently applied: all the matches obtained must be consistent with the order and orientation of the two flanking matches that were used to define the region.

Finally, to further increase specificity and alignment coverage, BLASTZ first aligns non-repeat sequences and then the matches are extended in repeat sequences. An additional 'trick' was used in the human–mouse alignment — both genomes were stripped of the lineage-specific repeats that have accumulated in both species since their divergence from their common ancestor. This enabled an increased alignment coverage from 32% to 40% of the human genome[7,45]. The default scoring matrix in BLASTZ was derived from previous alignments of known human–mouse homology regions[46] and was therefore optimized to give its best performance for the human–mouse comparison (although it has also been used successfully in human–*Fugu* localized alignments[37]). However, a customized scoring matrix can be provided by the user, if desired.

As a local aligner, BLASTZ detects PARALOGOUS and ORTHOLOGOUS similarities. However, in the human–mouse BLASTZ alignment, only a small proportion of the human genome (3.3%) aligned to multiple regions of the mouse genome. The resulting overlapping matches need to be carefully examined to sort out the orthologous alignment. Hence, a post-processing step is needed that involves selecting the most significant matches. Overall, the final alignment covered 98% of the known coding regions in the mouse and human genome, indicating that BLASTZ is highly sensitive for identifying well-conserved coding regions in known genes.

BLAT[47] is another local aligner that can be used in untranslated and translated modes. The untranslated mode was designed to align cDNA to genomic sequences of the same species. It is less effective at <90% sequence identity. The translated mode is more effective for comparative analysis. Using sequences that are masked for repeats and low-complexity regions rather than unmasked sequences, it produces a cleaner and faster output. The translated mode was initially inspired by the ExoFish method[36], which used the classical TBLASTX, with a particular identity matrix and parameters, to specifically align the coding regions that are conserved between human and *Tetraodon* (a freshwater pufferfish). This mode uses steps 1 and 2 (FIG. 3), so it produces a set of ungapped alignments, and gains in processing speed at the expense of overall sensitivity. Another limitation is that using the translated mode will more efficiently identify the regions that are conserved for their coding ability rather than for their regulatory functions. The seeding in BLAT is achieved by building an index of non-overlapping k-mers (in the translated mode, usually 5-mers) in one of the genomes. However, highly frequent k-mers and ambiguous sequences are excluded from the index. Although the index takes 30 minutes to generate, it can be stored and reused. In the next step, thousands of query sequences (<200 kb is recommended for each query sequence) from the other genome are rapidly scanned through the index. To obtain the final set of maximum-scoring ungapped matches, a DP alignment stage is applied upstream and downstream of each seed.

The global alignment can be divided into three steps. Step one involves the identification of the maximal repeated region between the two sequences that are being considered. The resulting pairs of exact-matching sequences that are flanked by mismatches are not necessarily unique and could overlap each other. They can be divided into clean matches (outside of known repeats) and repeat matches. Step two identifies a set of anchors. In this step, non-overlapping non-cross-matching matches are selected, with clean matches considered first, followed by repeat matches. For a set of n anchors, there will be n+1 unaligned regions between these anchors. To align those regions, step one (using the already processed list of matches) and step two are recursively applied for each interanchor pair that is defined in the preceding recursion step, until either no bases remain to be aligned or no significant matches can be found. In step three, for those sequences that remain to be aligned and that are <4 kb, a classical NW algorithm is applied. If sequences are >4 kb, the lack of anchors indicates that no significant alignment could be obtained between them.

AVID[48,49] is a global aligner, which assumes that the two sequences that are to be aligned are strictly homologous and have not undergone changes such as gene duplication, inversion or translocation, during evolution — in other words, it assumes that the order of the functionally conserved region has been preserved. This assumption is incorrect when applied directly to a whole genome; so, to achieve a whole-genome alignment, AVID needs a pre-processing step to identify the putative homologous unrearranged sequences or SYNTENIC REGIONS. Because of its speed and sensitivity, BLAT can be used for this purpose. BLAT hits are further filtered for specificity and are considered as putative SYNTENIC ANCHORS, which are then grouped if found on both genomes in the same order, the same orientation and separated by <250 kb (the sequence size used by AVID human–mouse comparisons). For groups <65 kb apart, the regions are extended 5′ and 3′ by ≥50 kb, depending on the group size. Each group defines the regions to be compared by AVID. The pre-processing step, therefore, results in a list of global alignments to be made.

Like AVID, LAGAN[50] is a global aligner that uses the same three-step strategy, with recursion of steps one and two. The advantages of LAGAN over AVID are that it can align larger sequences (>2 Mb), owing to the considerably lower memory requirements, and that it uses a different matching algorithm in step one — AVID looks for exact matches, limiting the comparisons to closely related organisms, whereas LAGAN was designed to align both distantly and closely related organisms. It uses the highly sensitive CHAOS program (which implements step one and two) that detects local alignments

PARALOGUES
Sequences, or genes, that have originated from a common ancestral sequence, or gene, by a duplication event.

ORTHOLOGUES
Sequences, or genes, that have originated from a common ancestral sequence, or gene, by a speciation event.

SYNTENIC REGION
A genomic region that is collinear in the order of genes (or of other DNA sequences) in a chromosomal region of two species.

SYNTENIC ANCHORS
Short aligned segments between genome sequences from two species, which are believed to define an orthologous relationship.

© 2003 Nature Publishing Group

Box 2 | **Dynamic programming**

Dynamic programming (DP) is a method for breaking down the alignment of sequences into small parts, which considers all the possible changes in moving from one pair of characters in the alignment to the next. It is comparable to moving across a DOT PLOT MATRIX **and keeping track of all the matching pairs, adding up those pairs that are found along a diagonal and making a subtraction when insertions are necessary to maintain an alignment. In this simple example, a match is scored as 1, a mismatch is scored as –1 and an insertion/deletion gap penalty is scored as –1.**

Example 1

	A	C	G	T	A
A	1	0	–1	–2	–3
C	0	2	1	0	–1
G	–1	1	3	2	1
T	–2	0	2	4	3
A	–3	–1	1	3	5

Example 2

	A	C	G	T	A
A	1	0	–1	–2	–3
C	0	2	1	0	–1
G	–1	1	3	2	1
C	–2	0	2	2	1
T	–3	–1	1	3	2
A	–4	–2	0	2	4

Example 1

A C G T A
| | | | |
A C G T A

Example 2

A C G C T A
| | | | | |
A C G – T A

Alignments are easily produced by finding the highest score, then tracing back through the matrix that has previously been defined.

using short inexact words (the level of degeneracy can be modified by the user). If used in the translated mode, CHAOS works best in conserved coding regions. So far, LAGAN has been applied to align rat–human and rat–mouse comparisons (see online links) in conjunction with BLAT, in the same way as BLAT/AVID.

MLAGAN — an extension to LAGAN — enables the multiple alignment of large genomic sequences. It involves a progressive alignment phase, based on LAGAN, which first aligns the genomes of the most closely related organisms, then incorporates the others in order of phylogenetic distance (this information is provided by the user). For example, when aligning human, mouse and rat, mouse and rat will be aligned first, followed by human. The application of multiple alignment to several whole-genomes has not previously been published. However, as the genome sequences of organisms such as rat, mouse and human are now available, all possible combinations of comparisons between them can be performed, and graphical interfaces such as mVISTA[28] or mPIPmaker[51,52] can be used to display the results, taking one species as a reference. Whole-genome multiple alignment seems to be the next challenge.

None of the programs mentioned so far (with the exception of translated BLAT and CHAOS) take into account the degeneracy of the genetic code in the final alignment stage — different triplets that code for the same amino acid cannot be aligned with a good score. WABA has been implemented to address this issue[38]. WABA aligns sequences in three phases. The first phase is fast and is similar to the seeding step of BLAT, but it is based on nucleotides and uses the two weighted-spaced seed 6of8 (11011011; BOX 3), in which the third position of a triplet is allowed to mismatch. No extension step is applied to these seeds, but they are grouped to define homologous regions, imposing the same order, orientation and distance constraints between them. As in BLAT, isolated matches that correspond to potential repeat sequences of period 4 or less (such as AAAAAAAA, CACACACA and ACGTACGT) or that contain ambiguous nucleotides are eliminated. The second phase aligns the homologous regions (2 kb against 5 kb) defined in phase one, using a pair-HIDDEN MARKOV MODEL (pair-HMM) method[53] (see the gene-prediction section for a detailed discussion) that, while aligning, can identify regions that are highly conserved (~90% base similarity) or lightly conserved (~50% base similarity) and apply a different scoring scheme, allowing for large insertions without giving them too high a penalty. The pair-HMM method is a local aligner that does not model inversion and duplication, and expects to align unrearranged sequences. As in the first phase, the HMM allows mismatches in the third position of a triplet. The third phase joins overlapping matches from the second phase. The pair-HMM method makes WABA slower than the other programs previously mentioned, but it is still rapid enough to have been applied to compare the *C. elegans* and *C. briggsae* genomes (which are each 100 Mb).

As its HMM has been optimized for these two organisms, WABA is not immediately suitable for other species comparisons. It is, however, an elegant method and was the first implementation of the pair-HMM method, which has become an increasingly important concept in gene-prediction programs that use comparative genomics.

Biological correctness. So, which one of these methods is best? Unfortunately, there is no clear answer to this question. First, our theoretical understanding of what is happening in evolution over time is insufficient to indicate which approach is superior. Second, different algorithms are tuned to different species comparisons (for example, the BLASTZ scoring scheme to mouse–human alignments[46] and WABA to *C. elegans*–*C. briggsae* alignments). Third, programs are developed and applied to answer different biological questions, for example, to simultaneously align as much of the DNA as possible, regardless of whether it is under selection or not (AVID, LAGAN and BLASTZ) or to identify highly and moderately-conserved coding regions that are thought to be under selection (BLAT and WABA).

It is important to decide whether we are interested only in the homologous base pairs, or in aligning homologous base pairs that have been conserved for a function. For biologists, the latter is more relevant, but most algorithms inherently try to maximize the former. For example, in the mouse–human alignments, ~40% of the base pairs are alignable, whereas only ~6% are estimated to be under selection[7]. Furthermore, considering ancient repeats or the fourfold degenerate codons that are thought to be under neutral selection, it seems that the substitution rate varies across the genome[7]. This means that some regions in the genome seem to have undergone more mutations and have evolved more rapidly than others.

DOT PLOT MATRIX
A visualization technique that allows the easy identification of matching nucleotides or amino acids (letters) between two sequences. For example, for two sequences X and Y, each letter has a unique coordinate on the x axis and the y axis respectively. When two letters are the same at a specified coordinate, a dot is plotted in the matrix at that position.

HIDDEN MARKOV MODEL (HMM). A probabilistic model that is applied to protein- and DNA-sequence pattern recognition. HMMs represent a system as a set of discrete states and as transitions between those states. Each transition has an associated probability. HMMs are valuable because they enable a search or alignment algorithm to be built on firm probabilistic bases, and the parameters (transition probabilities) can be easily trained on a known data set.

© 2003 Nature Publishing Group

Chiaromonte *et al.* have shown that similar regions of variability exist in mammalian genomes[54], indicating that some regions of the genome are flexible and tolerate more point mutations and repeat insertions, whereas other regions are rigid and accumulate fewer changes. Such conservation could be observed in large (~27 kb) non-coding regions[54]. The first obvious explanation is that this conservation is the result of selection. For non-coding regions, selection could be a result of the presence of a large number of regulatory elements, and this argument was used to explain the small number of repeats in the human HOX gene clusters[3]. The alternative explanation is that some regions have undergone a small number of mutations (mutation 'cold' spots) and their conservation has nothing to do with selection. Therefore, users should not conclude that a particular region must have functional relevance just because it is aligned.

Gene prediction

Defining gene structure in metazoan genomes remains a challenge, owing to the poor signal-to-noise level of the elements that define individual genes. Comparison with an evolutionarily related genome sequence should provide useful additional information for gene-structure definition. The so-called 'dual genome' gene predictors have great promise for the definition of metazoan genes. As with whole-genome alignment, the theory and practical requirements of running these programs rapidly becomes complex, but a basic understanding of the different assumptions that these programs make can help users understand their strengths and limitations. For many users, the precomputed datasets that are available at the Ensembl and UCSC web sites are sufficient, but, for individual cases, users with their own datasets might want to investigate running the programs through the web themselves.

Dual genome-prediction tools attempt to model a specific type of negative selection; the assumption is that when two sequences are aligned, most of the observed differences are neutral, and areas that are devoid of mutations are conserved because the mutations were detrimental to the organism. The information from modelling these detrimental mutations can be combined with our understanding of splicing, to provide better gene models. For example, for an amino acid such as glycine, which is encoded by four codons that all have the same first two base pairs, we expect mutations in the final base pair to be unrestrained, whereas mutations in the first two base pairs (which change the amino acid) are selected against.

Dual genome predictors can be subdivided into three classes: pair-HMM approaches that take a rigorously mathematical approach to the determination of joint gene structure and alignment; informant approaches that generally fix on the alignment and use the additional information to provide better gene prediction; and exon-finding approaches that attempt to demark the exons without trying to splice them together. Single-gene prediction has also used other techniques, such as linear DISCRIMINANT FUNCTIONS or NEURAL NETWORKS[55], but these have yet to be applied to comparative gene findings.

Pair-HMM methods. A group of methods based around the well-developed theory of HMMs are used by bioinformaticians to solve other sequence-related problems, such as gene-structure prediction on a single genome[56]. The pair-HMM method considers there to be one HMM that can generate two paired sequences, each of which contains a set of orthologous genes. Codons are modelled as either being aligned between the two sequences or independently inserted (representing gaps in one sequence relative to the other). Given two sequences, the pair-HMM method finds the most likely path to have generated these sequences, and provides both the alignment of the two sequences and the gene prediction inside that alignment.

Two full pair-HMM approaches have been developed — SLAM[57,58] and DoubleScan[59]. Both have to overcome two main problems: first, the parameterization of the HMM — that is, the selection of optimal parameters for a particular species combination given a representative set of genes — and second, the efficient decoding of the joint alignment and gene model. SLAM uses a generalized HMM approach for parameterization. DoubleScan uses an interesting parameterization that attempts to balance the coding and non-coding identity. Its creators claim that it should provide a more

Box 3 | **Two weighted-spaced versus consecutive-model seeding strategies**

Each alignment program uses a seeding strategy to identify where to start an alignment. The most common approach is to search for exact matching words (the consecutive model). BLASTN[44] uses this approach to identify matching words that are identical between the query sequence and the database sequence, using 11 consecutive identical nucleotides. For convenience, we denote the model as a string of 1 or 0, where 1 means identical base and 0 'does not matter'. So, the 11-consecutive model is noted 11111111111. The consecutive model allows a certain level of degeneracy to identify nearly exact words, as in the CHAOS program[50]. For example, a four-consecutive model with one level of degeneracy will identify words that match 1111, but also 1101 and 1011. The two weighted-spaced or non-consecutive model initially used in PatternHunter[41] and now in BLASTZ[45], specifies a single matching model in which the position for strict matches and mismatches that are allowed are strictly specified. In the case of BLASTZ, a 12of19 two weighted-spaced seed is used (111010011001010111). This precise combination was shown[41] to be the most sensitive among all combinations of the 12of19 two weighted-spaced seeds. It also shows a greater sensitivity than the 11-consecutive model if sequence similarity is >60%. Below this level of similarity, both models show the same sensitivity.

DISCRIMINANT FUNCTIONS
Classical statistical pattern-recognition methods that are used to categorize samples into two classes of data.

NEURAL NETWORKS
Mathematical models inspired by analogy with biological neurons to distinguish two or more classes of data.

© 2003 Nature Publishing Group

SCORE MAXIMIZATION PROCESS Many algorithms attempt to find the solution, under a scoring scheme, that is believed to best reflect reality. For 'simple' models, including hidden Markov models, precise mathematical formulae can be used that will guarantee to find the highest score.

robust parameterization, as the overall level of similarity between the two sequences should not alter the switching between coding and non-coding states; instead, only the pattern of substitutions should influence the decoding. To restrict the huge number of possibilities considered by pair-HMM algorithms, both methods use rough alignments: SLAM uses the AVID system to seed its regions, whereas DoubleScan uses BLAST.

Pair HMM methods are acknowledged to have a strong theoretical basis. However, at present they are no better than other methods (such as informant methods, described below), they are costly to run and they require two long orthologous sequences as their input.

Informant approaches. The informant-sequence approach also uses either explicit HMM parameterization or a SCORE MAXIMIZATION PROCESS. However, unlike pair-HMMs, informant methods use only one sequence to predict a gene structure; the other sequence provides information through its alignment, which is often called the informant alignment. Because of this asymmetrical approach, informant methods can use more varied inputs, even including unassembled reads aligned to a reference genome. Three main informant methods are currently available: Twinscan[60], SGP-2 (REF. 61) and GenomeScan[62]. Twinscan differs from the other two as it uses a base-by-base model of the informant sequence. GenomeScan and SGP-2 both use a more feature-based approach, in which the informant alignments are used to influence the score of exons. In all three cases, the parameterization of the interplay between the informant alignment and the gene prediction has to be precise; for this reason, these approaches generally run their own alignment method (often BLAST). Hence, the user can view them mainly as methods for predicting genes from a single sequence with help from another informant source of data.

Exon prediction. There is only one single-exon prediction system — ExoFish — that was designed to provide highly specific exon prediction from *Tetraodon* sequences[36]. ExoFish is a carefully parameterized version of TBLASTX, which sacrifices a certain amount of sensitivity (that is, coverage of exons) for high specificity; so, users can have reasonable confidence that the exon is true. ExoFish therefore provides an excellent starting point for designing PCR primers to get from a single exon to the complete gene structure with the help of *in vitro* experiments.

Which method to use? All these methods have had several successes over the past couple of years. Before the mouse genome sequence was available, most people had mistakenly assumed that comparative genomics would ensure an almost trivial decoding of the gene structure of many genes. In practice, those who design gene-prediction methods have had to work hard to provide a useful contribution over and above the action of *ab initio* predictors. A particular success was the work by Guigo and colleagues (R. Guigo, personal communication), who used informant methods combined with several simple criteria — in particular, the presence of a conserved intron — to produce a strong prediction of the number of common genes in mammalian genomes. This number is surprisingly low (25,000–27,000). It is interesting to note that the pair-HMM methods implicitly include the conserved intron as part of their model, and so, in theory, they should provide a more rigorous way of incorporating this extra information into gene finding. This and other studies have shown that the sensitivity and specificity of comparative gene finding is likely to increase with increasing numbers of genomic sequences.

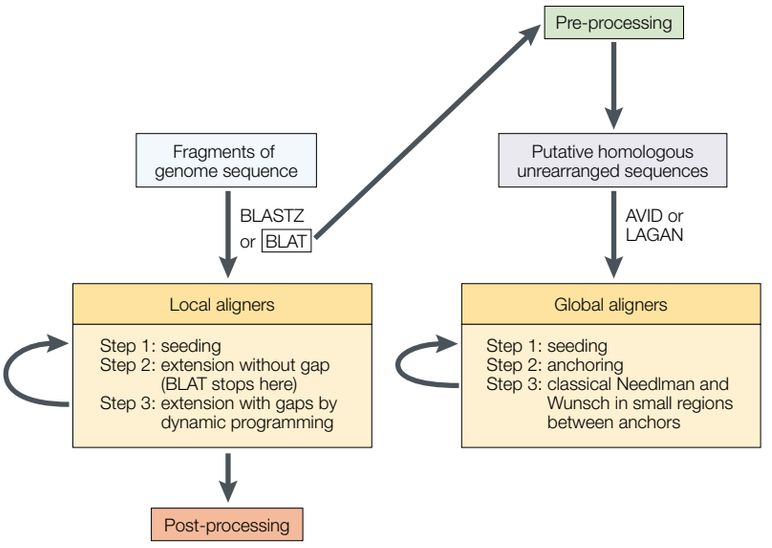


Figure 3 | **Schematic description of whole-genome alignment processes.** Both local (BLAT or BLASTZ) and global (AVID or LAGAN) aligners have three main steps in which the strategies that are used differ. In step 1, different seeding strategies are used. In step 2, the seeds can be used as nucleation points for gap-free extension or as anchor points to partition the alignment problem. In step 3, the dynamic programming process (Box 2) produces gapped alignments. Repetitions of the three steps are applied (looping arrows) — although there is only one repetition in BLASTZ — on non-aligned sequences until no sequences remain to be aligned, or they are too divergent to obtain a specific alignment. Local aligners can directly compare two genomes in an 'all-versus-all' approach: they align properly rearranged sequences and identify orthologous, as well as paralogous, similarities. A post-processing step is then used to sort the orthologous alignments. Global aligners assume that the two sequences that are to be aligned are strictly homologous — that is, that the order of the conserved region has been maintained. As a pre-processing step, the results from a local aligner (BLAT) are used to identify such homologous unrearranged sequences.

Finding regulatory regions

The discovery of regulatory regions in intergenic DNA through cross-species comparison is often termed phylogenetic footprinting (by analogy with DNAase footprinting)[63]. It is based on the observation that functionally important regions tend to have a lower mutation rate than non-functional regions. Indeed, several studies[34,64,65] have shown that putative transcription-factor binding sites are enriched in conserved non-coding genomic sequences (footprints). Many other studies have linked evolutionarily conserved regions to experimentally determined regulatory elements[66].

© 2003 Nature Publishing Group

Two main approaches have been used in the discovery of *cis*-regulatory motifs. The first approach tries to find common motifs that are shared by multiple orthologous sequences, whereas the second is based on the global alignment of orthologous sequences, followed by the identification of conserved regions. In both cases, additional information, such as previously known motifs, can be used to enhance the signal.

Which region to use? To consider all possible *cis*-acting elements, it is necessary to analyse the non-repetitive intergenic sequences of the whole regulatory region of true homologous genes. In theory, for metazoan genes, the whole regulatory region should include the 5′- and 3′-flanking regions of the genes, as well as the intronic sequences. For a human gene of ~30 kb, this is a considerable amount of sequence. Analysis is greatly hampered by two additional features. First, the 5′ end of genes is often the least well-defined area, as it is necessary to have experimental evidence of the promoter that stretches to the transcriptional start point. Second, there are known enhancers that work several kilobases away from the transcriptional start, so the boundaries of the regions to be used are unclear.

Most successful examinations, therefore, have used both careful manual analysis of the available experimental evidence to define the true start of the gene, and a mixture of educated guessing and systematic comparisons to find potential *cis*-regulatory elements[67]. In more densely packed genomes, such as those of *C. elegans* or *D. melanogaster*, neighbouring genes have been used as the limits of the amount of DNA that is considered[68]. Dieterich *et al.*[69] used the experimentally defined promoter regions in the eukaryotic promoter database (EPD)[70] to estimate the distribution of the locations of mammalian promoters with respect to their genes. They concluded that 10 kb upstream from the start of translation is sufficient in most cases. Nonetheless, the criteria for the selection of the region to be examined are still poorly defined.

Because genome annotation and untranslated region (UTR) boundaries are not perfect, it is important to ensure that the analysed sequences do not contain unannotated coding regions, or are not constrained by mRNA structure. In addition, researchers have been systematically removing repeats from the analysis using conventional tools like RepeatMasker, but, in some cases, conserved repeat elements, such as Alus, have been shown to be involved in gene expression[71].

Evolutionary issues. What are the evolutionary relationships between genes and their *cis*-regulatory motifs? For example, two genes might be strictly orthologous but, in fact, their regulation has diverged because the two species have different uses for the orthologues. More significant is the case of paralogous families, in which the largest changes in function are expected to occur as a result of changes in expression patterns[72]. Indeed, according to Force *et al.*[73], one mechanism by which duplicated genes might be retained is the evolution of new expression sites through changes in the regulatory-control elements of both copies. An example of the divergence of two genes being driven by changes to expression (and hence, in evolutionary terminology, being under positive selection) are the *engrailed* genes in zebrafish[73]. Establishing true orthologous relationships is therefore important, especially between remote species. When comparing mouse to human, Dermitzakis and colleagues[74] paid particular attention to confirming that the compared genes were homologous, by combining a BLAST search with gene annotation in the National Center for Biotechnology Information (NCBI), Genbank and Mouse Genome Informatics (MGI) databases to ensure that the study of their *cis*-regulation made sense.

The question of how evolution affects *cis*-regulatory motifs is so poorly understood that no formal models for this process have been proposed (in contrast with the evolution of protein sequences, which is underpinned by much stronger formal models). This means that it is up to the intuition of researchers as to which species should be used in the comparison. Ideally, the orthologous genes studied would have retained the same expression pattern in different species that are sufficiently evolutionarily distant that only functionally conserved sequences are alignable. However, there seems to be no ideal comparison to make: intra-mammal comparisons (such as mouse–human) show a large amount of non-functional conservation[7], but in intravertebrate analyses (such as fish–human) the signal is hard to detect, and it is less clear whether functional conservation should be expected between the same motifs. More worrying is that NEUTRAL DRIFT of the sequence could 'create' and 'destroy' *cis*-regulatory sites at an appreciable rate — a phenomenon that does not have a large role in protein evolution. For example, Ludwig *et al.*[75] showed that the *even-skipped* stripe-two element, which controls the expression of *even-skipped* stripe-two mRNA, had undergone substitutions and deletions that modified the *cis*-regulatory sites and their spacing in two species as closely related as *D. melanogaster* and *D. pseudoobscura*. Nevertheless, the expression of *even-skipped* stripe two remains strongly conserved between the two species, raising the possibility of compensatory mutations. Ludwig *et al.* suggest that STABILIZING SELECTION has enabled the mutational turnover of functionally important sites. They predict that such a pattern of substitution will be common in *cis*–regulatory regions.

With more sequences available, and more interest in this area, we expect there to be significant gains in our understanding of how evolution affects *cis*-regulatory motifs. An important area of research is, therefore, the identification and modelling of the variety of evolutionary processes that shape the structure of DNA. Recently, this issue was addressed by trying to distinguish regulatory regions from neutrally evolving DNA by using genomic-sequence alignments[76]. The authors show that a Markov model procedure that exploits

NEUTRAL DRIFT
The process by which a DNA sequence acquires many mutations over time that have no phenotypic effect, and are not acted on by Darwinian selection.

STABILIZING SELECTION
Selection that favours intermediate phenotypes over extreme phenotypes.

© 2003 Nature Publishing Group

Table 1 | **Web-based programs for genomic alignment and the retrieval of potential transcription-factor binding sites**

Program	Alignment program	TFBS database	Program used to locate TFBS
Consite	DPB	Consite database non-redundant experimental binding-site studies	Consite
rVISTA	AVID	Transfac/user defined	Match
Toucan	AVID	Transfac	Motif scanner
Trafac	BLASTZ	Transfac	MatInspector or Match

TFBS, transcription-factor binding site.

short-pairing patterns successfully discriminates between regulatory regions and neutral features.

Motif overrepresentation. Motif-finding techniques are used to find overrepresented motifs in a cluster of genes that have, for example, been selected on the basis of expression analysis. MEME[77], AlignACE[78] and DIALIGN[79] are three examples of such techniques, but many more have been reported in the literature[80–82]. Although these approaches can theoretically be used for phylogenetic footprinting, they were designed to compare independent sequences and do not take into account the phylogenetic relationship between homologous sequences. So, the basic assumption that the underlying sequences share only the motifs of interest is not valid.

To overcome this problem, Blanchette and Tompa[83] developed another method — Footprinter — which is more suitable for comparing orthologues, and identifies all the DNA motifs that have evolved at a slower rate than the surrounding sequence. The input data are a list of orthologous sequences and the phylogenetic relationships between these sequences or the corresponding species. The authors have successfully used this program on nine different examples for which multiple orthologous sequences and known regulatory regions were available.

An important drawback of this approach is its relatively bad performance on a small set of orthologues. Indeed, all motif-finding techniques work better with increasing amounts of sequence, as the distinction between the conserved motifs and the diverged background becomes clearer. Blanchette and Tompa[83] systematically used more than three species and increased the number of sequences by including paralogues. With the sequencing of further genomes this will become less of a problem, but the approach will only work if most of the genomes retain sufficiently conserved motifs within the analysed segment; this as discussed in the previous section, is probably not normally the case. Furthermore, these methods work less well with large sequences, and as metazoan promoters might lie a considerable distance away from the transcription start site, their usefulness might be limited.

Nevertheless, these approaches can find motifs that independently satisfy the initial parameters of the surrounding sequence identity. This is not the case with global alignment, in which the noise of the diverged non-functional background can overcome the short conserved signal.

Alignment. Alignment techniques have been used extensively to discover *cis*-regulatory elements. This approach consists of aligning regions of homology in the non-coding sequences in the vicinity of orthologous genes from two or more species. Most of the work has been done on well-studied examples like the α-globin cluster[67], the SCL loci[84], the *Oxb4* gene[66] and others[85,86] — regions that often correspond to loci that are involved in human diseases. Nevertheless, more general analyses have been performed on whole genomes[64,68,76] and their functional subsets.

The computational algorithms used in such comparative sequence analysis are based either on local or global alignments. Because regulatory elements tend to be short conserved sequences, algorithms like DBA[35] or BBA[87], which focus on aligning highly conserved blocks while also allowing large gaps, are theoretically better suited for identifying *cis*-regulatory regions. In practice, any alignment techniques will pick up the regulatory elements that are located in 'modules' of long highly conserved regions. But how many regulatory elements are located in non-conserved segments? An increasing number of studies show evidence of the modular organization of *cis*-regulatory sites[88] but other studies have shown examples of regulatory elements that are found in regions of low sequence identity. For example, Flint *et al.*[67] reported the presence of a pair of functional MARE motifs in pufferfish and chicken that is located in a poorly conserved region compared with human and mouse.

The aligning programs used are those described in the previous section on aligning genomes. In addition, because the amount of sequence to align is relatively small in the case of regulatory regions, other algorithms such as ClustalW, BLAST and the Waterman-Eggert algorithm can be used to obtain better alignments[69]. In their global comparison of the human and mouse genome sequences, Levy and Hannenhalli[89] used, as a first step, unique anchors of 40 bp with ≤80% homology extending 20 kb upstream or downstream. The resulting mouse and human sequence pairs were subsequently realigned using BLAST.

Substantial resolving power is added by including more than two sequences in a multiple sequence alignment, as each lineage diverged independently after separation from a common ancestor. Examples of programs that are used in such alignments include Yama2 (REF. 90), ClustalW[91], Multalign[92] and DIALIGN[79]. Because DIALIGN does not have a GAP PENALTY and starts by identifying short conserved regions, this algorithm is more suitable for identifying regulatory regions than ClustalW.

GAP PENALTY
Alignment programs deal with insertions and deletions (indels) by introducing a 'gap' in the sequence that contains the deletion. The introduction of gaps and their extension decreases the overall alignment score by a certain value. This value is defined by a gap-opening penalty and a gap-extension penalty, both of which are used as parameters in alignment programs.

© 2003 Nature Publishing Group

Once the alignment has been made, conserved sequences need to be located. For alignments of only two sequences, a simple heuristic 'X % conservation over at least Y nucleotides' is generally used. Dubchak *et al.*[86] used two alignments (human–dog and human–mouse) and a 200 kb sequence from human chromosome 5q31 to define the cut-off criteria X and Y for conserved sequences based on maximizing the percentage of the regions that are common in three species. In other cases, a simple ranking of alignment scores seems to give better results[67] compared with fixed settings. For multiple alignments, more parameters must be taken into account, such as the phylogenetic relationship between species and the nucleotide frequencies at each position. Stojanovic *et al.*[93] developed and compared several methods for finding these conserved blocks on multiple sequences. These methods take into account the level of similarity for each aligned column, adding parameters such as the overall nucleotide composition, phylogenetic relationship and reference sequence. Stojanovic *et al.*[93] found that although these methods have different theoretical underpinnings, none seemed clearly superior to the others as long as the parameters that had been chosen were optimal.

Because global alignments only provide information about which region is common to two or more species, the challenge for these techniques is to assess whether these regions of homology are involved in regulation. This is why global alignment has often been used in conjunction with known transcription-factor binding sites, usually taken from the Transfac database[94] — a database of transcription factors and their binding sites. Match or MatInspector[95], which are software tools that use a library of matrix descriptions to locate matches, are used together with the knowledge of functional conservation to predict binding sites. Trafac[96], rVISTA[97], Toucan and Consite are programs that are available on the web and enable the user to create alignments and visualize the results (TABLE 1).

Conclusion

Studying the effect of evolution on specific sequences and functions has long been a fruitful approach to elucidate biological mechanisms. With the advent of whole-genome sequences, this study has now been extended across the entire genome of several organisms. The technology to effectively work with large vertebrate-sized genomes has just started to develop, and we can expect more genomes to be sequenced and more techniques to be produced. Users who want to focus at the biological level should probably choose the precomputed runs on web sites that are likely to be organized around the whole-genome displays at Ensembl, UCSC and NCBI. Other users might want to examine genome regions in more detail, and use the web-based tools, the existing programs, or, indeed, develop their own.

It is likely that the long-standing theories of the evolution of sequences will have a stronger role in our interpretation of this comparative data. There are three main advances to be made in the future. First, an increased number of genomes will improve the analytical power of all methods. For example, we believe that by the fourth or fifth sequenced mammalian genome, comparative gene finding will provide a strong computational signal for the common set of genes across all mammals. Similarly, power will be added by other vertebrate genomes, although only for the gene set that is common to all vertebrates. Second, increased power is likely to be provided in the area of biological models of how negative selection works for different functional constraints. At the moment, the protein-coding genes are best understood, but we believe that increased efforts will be directed towards other areas — in particular, towards the regions that control gene expression. Third, from the theoretical perspective it is likely that more effort will be placed on modelling positive selection, that is, on identifying the mutations that improve the existing function or create a new one. Overall, these advances will make the use of comparative genomic data as widespread as the use of primary genome data.

1. The *C. elegans* Genome Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
2. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
5. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
6. Holt, R. A. *et al.* The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149 (2002).
7. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
The first whole-genome comparative analysis of two mammalian organisms.
8. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157–2167 (2002).
References 1–8 are the original publications for some of the sequenced 'entire' metazoan genomes.
9. Bolshakov, V. N. *et al.* A comparative genomic analysis of two distant diptera, the fruit fly, *Drosophila melanogaster*, and the malaria mosquito, *Anopheles gambiae*. *Genome Res.* **12**, 57–66 (2002).
10. Zdobnov, E. M. *et al.* Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159 (2002).
One of three papers about the *Anopheles* genome, which gives a good flavour of more distant comparative genomics compared with the inter-mammal papers.
11. Dehal, P. *et al.* Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293**, 104–111 (2001).
12. Mural, R. J. *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661–1671 (2002).
13. Hedges, S. B. The origin and evolution of model organisms. *Nature Rev. Genet.* **3**, 838–849 (2002).
14. Graur, D. & Wen-Hsiung, L. *Fundamentals of Molecular Evolution* (Sinauer Associates, Inc., Sunderland, Massachusetts, 2000).
15. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).
16. Freeling, M. Grasses as a single genetic system: reassessment 2001. *Plant Physiol.* **125**, 1191–1197 (2001).
17. Van Sluys, M. A. *et al.* Comparative genomic analysis of plant-associated bacteria. *Annu. Rev. Phytopath.* **40**, 169–189 (2002).
18. Edwards, R. A., Olsen, G. J. & Maloy, S. R. Comparative genomics of closely related salmonellae. *Trends Microbiol.* **10**, 94–99 (2002).
19. Brosch, R., Pym, A. S., Gordon, S. V. & Cole, S. T. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol.* **9**, 452–458 (2001).
20. Paulsen, I. T., Chen, J., Nelson, K. E. & Saier, M. H. Comparative genomics of microbial drug efflux systems. *J. Mol. Microbiol. Biotech.* **3**, 145–150 (2001).
21. McClelland, M. *et al.* Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *Salmonella enterica* serovars, *Typhimurium*, *Typhi* and *Paratyphi*. *Nucleic Acids Res.* **28**, 4974–4986 (2000).
22. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
23. King, J. L. & Jukes, T. H. Non-Darwinian evolution. *Science* **164**, 788–798 (1969).
24. Ohta, T. & Tachida, H. Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. *Genetics* **126**, 219–229 (1990).
25. Miller, W. Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* **17**, 391–397 (2001).
26. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
27. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
28. Mayor, C. *et al.* VISTA: visualizing global DNA sequence

© 2003 Nature Publishing Group

alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047 (2000).
29. Harris, T. W. *et al.* WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.* **31**, 133–137 (2003).
30. Clamp, M. *et al.* Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31**, 38–42 (2003).
31. Karolchik, D. *et al.* The UCSC genome browser database. *Nucleic Acids Res.* **31**, 51–54 (2003).
32. Giardine, B. *et al.* GALA: a database for genomic sequence alignments and annotations. *Genome Res*. (in the press).
33. Pennacchio, L. A. *et al.* An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294**, 169–173 (2001).
34. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human–mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26**, 225–228 (2000).
35. Jareborg, N., Birney, E. & Durbin, R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9**, 815–824 (1999).
36. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
The first large-scale comparison between two vertebrate genomes.
37. Gilligan, P., Brenner, S. & Venkatesh, B. *Fugu* and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene* **294**, 35–44 (2002).
38. Kent, W. J. & Zahler, A. M. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. *Genome Res.* **10**, 1115–1125 (2000).
The first software implementation of a pair-HMM to align sequences.
39. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999).
40. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).
41. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).
The publication that proposed the original two weighted-spaced model to identify nearly exact matching words
42. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**, 950–958 (2000).
43. Mullikin, J. C. & Ning, Z. The phusion assembler. *Genome Res.* **13**, 81–90 (2003).
44. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
45. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
46. Chiaromonte, F., Yap, V. B. & Miller, W. Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput*. 115–126 (2002).
47. Kent, W. J. BLAT — the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
48. Bray, N., Dubchak, I. & Pachter, L. AVID: a global alignment program. *Genome Res.* **13**, 97–102 (2003).
49. Couronne, O. *et al.* Strategies and tools for whole genome alignments. *Genome Res*. **13**, 73–80 (2003).
50. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*. (in the press).
51. Schwartz, S. *et al.* PipMaker — a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586 (2000).
52. Elnitski, L. *et al.* PipTools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. *Genomics* **80**, 681–690 (2002).
53. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. in *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* 80–99 (Cambridge Univ. Press, Cambridge, UK, 1998).
54. Chiaromonte, F. *et al.* Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proc. Natl Acad. Sci USA* **98**, 14503–14508 (2001).
55. Zhang, M. Q. Computational prediction of eukaryotic protein-coding genes. *Nature Rev. Genet* **3**, 698–709 (2002).
56. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
57. Alexandersson, M., Cawley, S. & Pachter, L. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res*. (in the press).
The most complete pair-HMM model for gene prediction that has been implemented so far.
58. Pachter, L., Alexandersson, M. & Cawley, S. Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comput. Biol.* **9**, 389–399 (2002).
59. Meyer, I. M. & Durbin, R. Comparative *ab initio* prediction of gene structures using pair-HMMs. *Bioinformatics* **18**, 1309–1318 (2002).
60. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17** (Suppl.) S140–S148 (2001).
This paper describes Twinscan, which is one of the best informant-HMM gene-modelling approaches.
61. Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T. & Guigo, R. SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res.* **11**, 1574–1583 (2001).
62. Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803–816 (2001).
63. Tagle, D. A. *et al.* Embryonic ε- and γ-globin genes of a prosimian primate (*Galago crassicaudatus*): nucleotide and amino-acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**, 439–455 (1988).
64. Levy, S., Hannenhalli, S. & Workman, C. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics* **17**, 871–877 (2001).
65. Fickett, J. W. & Wasserman, W. W. Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.* **11**, 19–24 (2000).
66. Aparicio, S. *et al.* Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. Natl Acad. Sci. USA* **92**, 1684–1688 (1995).
67. Flint, J. *et al.* Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α-globin cluster. *Hum. Mol. Genet.* **10**, 371–382 (2001).
68. Webb, C. T., Shabalina, S. A., Ogurtsov, A. Y. & Kondrashov, A. S. Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucleic Acids Res.* **30**, 1233–1239 (2002).
69. Dieterich, C. *et al.* Annotating regulatory DNA based on man–mouse genomic comparison. *Bioinformatics* **18** (Suppl.), S84–S90 (2002).
70. Praz, V., Perier, R., Bonnard, C. & Bucher, P. The eukaryotic promoter database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.* **30**, 322–324 (2002).
71. Hamdi, H. K., Nishio, H., Tavis, J., Zielinski, R. & Dugaiczyk, A. Alu-mediated phylogenetic novelties in gene regulation and development. *J. Mol. Biol.* **299**, 931–939 (2000).
72. Liu, T., Wu, J. & He, F. Evolution of *cis*-acting elements in 5′ flanking regions of vertebrate actin genes. *J. Mol. Evol.* **50**, 22–30 (2000).
73. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
74. Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**, 1114–1121 (2002).
The authors estimate that ~30–40% of the functional *cis*-acting elements in human are not functional in rodents.
75. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564–567 (2000).
A publication that studies the compensatory mutation and stabilizing selection of *cis*-acting elements in two species of *Drosophila*.
76. Elnitski, L. *et al.* Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13**, 64–72 (2003).
77. Bailey, T. L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 21–29 (1995).
78. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16**, 939–945 (1998).
79. Morgenstern, B., Frech, K., Dress, A. & Werner, T. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14**, 290–294 (1998).
80. Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).
81. Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.* **8**, 1202–1215 (1998).
82. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214 (2000).
83. Blanchette, M. & Tompa, M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12**, 739–748 (2002).
84. Gottgens, B. *et al.* Transcriptional regulation of the stem cell leukemia gene (*SCL*) — comparative analysis of five vertebrate *SCL* loci. *Genome Res.* **12**, 749–759 (2002).
85. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
86. Dubchak, I. *et al.* Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**, 1304–1306 (2000).
87. Zhu, J., Liu, J. S. & Lawrence, C. E. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14**, 25–39 (1998).
88. Berman, B. P. *et al.* Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA* **99**, 757–762 (2002).
89. Levy, S. & Hannenhalli, S. Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome* **13**, 510–514 (2002).
90. Chao, K. M., Hardison, R. C. & Miller, W. Recent developments in linear-space alignment methods: a survey. *J. Comput. Biol.* **1**, 271–291 (1994).
91. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
92. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881–10890 (1988).
93. Stojanovic, N. *et al.* Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.* **27**, 3899–3910 (1999).
94. Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316–319 (2000).
95. Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23**, 4878–4884 (1995).
96. Jegga, A. G. *et al.* Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.* **12**, 1408–1417 (2002).
97. Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I. & Rubin, E. M. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12**, 832–839 (2002).
98. Hedges, S. B. & Kumar, S. Genomics: vertebrate genomes compared. *Science* **297**, 1283–1285 (2002).
99. Venkatesh, B., Gilligan, P. & Brenner, S. *Fugu*: a compact vertebrate reference genome. *FEBS Lett.* **476**, 3–7 (2000).
100 Wittbrodt, J., Shima, A. & Schartl, M. *Medaka* — a model organism from the Far East. *Nature Rev. Genet*. **3**, 53–64 (2002).

Acknowledgements
We thank M. Brudno, W. Miller and L. Pachter for providing their respective manuscripts before publication and W. J. Kent, W. Miller, M. Brudno and L. Bentolila for helpful discussion and comments on the manuscript. We also thank the anonymous reviewers for many helpful suggestions. A.U.-V. is funded by the Wellcome Trust. E.B. and L.E. are funded by the European Molecular Biology Laboratory.

Online links

DATABASES
The following terms in this article are linked online to:
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
APOAV | *even skipped*

FURTHER INFORMATION
Berkeley Genome Pipeline: http://pipeline.lbl.gov
Comparative analysis of the rat genome: http://pipeline.lbl.gov/rat
ECR Browser: http://nemo.lbl.gov/ecrBrowser
Ensembl Genome Browser: http://www.ensembl.org
Penn State Bioinformatics Group: http://bio.cse.psu.edu
UCSC Genome Browser: http://genome.ucsc.edu
WormBase: http://www.wormbase.org
Access to this interactive links box is free online.

© 2003 Nature Publishing Group