Searching for regulatory elements in human noncoding sequences

Laurent Duret* and Philipp Bucher†

Important progress has been made in the past two years in the identification of Pol II promoters. For most other regulatory elements, however, current biological knowledge is still insufficient to allow the development of prediction tools. The phylogenetic-footprinting strategy, which is based on the comparative analysis of homologous sequences, is a very efficient approach to identify new unknown regulatory elements. The recent organization of large-scale sequencing projects for some model vertebrate organisms will be extremely valuable for the prediction of regulatory elements in the human genome.

Addresses
*Laboratoire de Biométrie, Génétique et Biologie des Populations, UMR 5558, CNRS Université Claude Bernard, 43 Boulevard du 11 Novembre 1918, F-69622 Villeurbanne cedex, France; e-mail: duret@biomserv.univ-lyon1.fr
†Institut Suisse de Recherche Expérimentale sur le Cancer, 155 chemin des Boveresses, CH-1066 Epalinges, Switzerland; e-mail: pbucher@isrec-sun1.unil.ch

Current Opinion in Structural Biology 1997, 7:399–406

http://biomednet.com/elecref/0959440X00700399

Abbreviations

bp	base pair
HCR	highly conserved region
HMM	hidden Markov model
kb	kilobases
LCR	locus control region
MAR	matrix attachment region
Mb	megabase
Myr	million year
nt	nucleotide
Pol	polymerase
SAR	scaffold attached region
TE	transcriptional element
UTR	untranslated region

Introduction

Large-scale genome sequencing projects currently determine hundreds of megabases (Mb) each year. Thus, one of the major challenges that biologists now have to take up is to extract relevant information from this huge amount of sequence data: to detect the genes embedded in these sequences and to try to decipher their function. The identification of regulatory elements required for the correct expression of genes is an essential step in the understanding of their function. This task is particularly arduous in the case of the large eukaryotic genomes, such as ours, that are replete with noncoding sequences, most of which are probably functionless. How can one identify a regulatory element overwhelmed in nonfunctional DNA? Obviously, the amount of experimental work that would be required to systematically analyze these noncoding sequences exceeds the ability of researchers. Hence, there is an urgent need for computational tools that identify potential regulatory elements with which researchers could focus their experiments. Furthermore, such tools may also be of practical interest by improving methods of gene recognition.

Two major classes of regulatory regions can be distinguished. The first class includes elements that are recognized at the DNA level: promoters; enhancers; locus control regions (LCRs); and matrix attachment regions/scaffold attached regions (MAR/SARs). The second class includes elements that are involved in post-transcriptional processes, and that are recognized at the RNA level: *cis*-acting elements responsible for the regulation of processing, transport, translation and stability of mRNAs.

Whereas many efforts have been made for locating protein-coding regions within genomic sequences (for a review, see [1]), or for predicting the structure and function of proteins (e.g. see Jones, pp 377–387, this issue), relatively few tools have been developed for the prediction of regulatory elements. Two main types of approaches can be distinguished. The first one includes methods that rely on biological knowledge to set up rules or strategies to predict regulatory elements. The most advanced field of such 'knowledge-based methods' is the prediction of RNA polymerase (Pol) II promoters. Recent advances in this domain are described in the first part of our review. The second type of approach relies on the comparative analysis of homologous sequences. Although such approaches are not new, they are now gaining importance thanks to the recent set-up of sequencing projects for several genomes of vertebrate model organisms. The second part of our review summarizes the theoretical background for this comparative approach and illustrates with recent examples its efficiency, both for the study of some particular genes and for large-scale genome sequence analysis.

Knowledge-based methods: prediction of vertebrate Pol II promoter sequences as a paradigm

RNA polymerase II promoters are arrays of regulatory elements (transcriptional elements [TEs]) that are relatively short sequence motifs (5–25 bp in length) and that are recognized by regulatory proteins (transcription factors). Current knowledge of these promoters relies on extensive experimental work: over 4000 transcriptional elements have been described, corresponding to binding sites of hundreds of transcription factors [2•]. Databases such as TRANSFAC [2•] or TFD [3] have been developed to

compile information relative to these TEs, their binding factors and their consensus sequences (n.b. TFD has not been updated since September 1993). This information allows the establishment of rules or strategies that can be used for promoter prediction.

Prediction of gene regulatory signals

Most transcription regulatory elements are highly degenerate sequence motifs that are recognized by DNA-binding proteins. The computational problem of defining such signals can be logically subdivided into three parts: first, the problem of mathematically representing a degenerate sequence motif; second, the problem of deriving a specific motif description from the available sequence and functional information; and third, the problem of searching for signal occurrences in new sequences once the signal is defined.

Until recently, most experimental biologists and many computational biologists have been using IUPAC code based consensus sequences as a means of describing a transcription factor binding site or a physiologically defined control signal. However, such descriptors have an important limitation: they cannot distinguish between mismatches of varying degrees of severity. More realistic representations of regulatory signals may be obtained by using weight matrices. The recently introduced hidden Markov models (HMMs; for a review, see [4•]) add an additional level of flexibility by allowing variable spacing between conserved blocks — a feature that may be essential for accurately describing the binding preferences of large multimeric transcription regulatory complexes.

Of course, a nearly infinite number of possible consensus sequences or weight matrices exist. Knowledge-based approaches to derive signal descriptions attempt to find the best one with regard to given sequence and functional data. Biological function may be related to sequences in two ways: first, via a physiological process that has been mapped to a specific site within a sequence; and second, via a biological or biochemical activity associated with a DNA fragment or a sequence region delimited by mutagenesis. In the first case, the goal of computational analysis is to identify sequence motifs that occur at characteristic distances from the physiological site. In the second case, one searches for a common motif within the functionally defined set of sequence regions. The standard technique for searching motifs that are positionally correlated with a site proceeds by analyzing the sequence contents in a sliding window along an ungapped sequence alignment (e.g. [5]). The most commonly used weight-matrix descriptions of the major constitutive elements of eukaryotic promoters have been derived by a heuristic variation of this general method using weight matrices instead of consensus sequences. More recently, neural networks [6] and HMMs [7•] have been applied to promoter sequences aligned by the transcription start sites.

The classical methods to find a common motif in a set of functionally related sequences proceed via a multiple segment alignment. The goal is to arrange the sequences in such a way that DNA bases interacting with the same molecular components of the transacting factor are superimposed. Note that this definition of a correct alignment differs from the one used in evolutionary studies. The segment-alignment problem is very difficult to solve if the sequences share only weak similarities, and the development of corresponding algorithms has therefore been a major research focus in the comparative analysis of DNA regulatory elements. The introduction of Expectation-Maximization algorithms and Gibbs-sampling techniques have been important steps forward [8]. Recent attempts to improve segment alignments have focused on the recognition of motifs in noisy input data [9,10•], and on the simultaneous recognition of multiple motifs in biochemically heterogeneous sequence sets [11•]. Although important progress has undoubtedly been made in this field recently, it deserves mentioning that an elegant but largely forgotten heuristic algorithm devised by Queen *et al.* [12] will in many instances produce alignments equivalent or even superior to those obtained by current standard methods.

Once a signal is defined, in the form of a consensus sequence or a weight matrix, the search for signal occurrences is relatively straightforward. The current state of the art software tools that search for transcription factor binding sites using libraries of weight matrices are MatInspector [13•] and MATRIX SEARCH [14•], which is now included in SIGNAL SCAN [15•]. Although these programs make the best use of current knowledge, it must be recognized that the accuracy of the prediction is a function of the quality of the weight matrices in the reference collections. Unfortunately, many of these matrices have been derived from small sequence sets, and thus cannot be expected to make reliable predictions.

New promoter prediction methods

Even if good progress has recently been made in the identification of putative TEs, such predictions are not sufficient to correctly recognize promoters. Because of the short and degenerate nature of TEs, many of the putative elements that are identified in any sequence scan are in fact biologically irrelevant. For example, when using position weight matrices of the TATA box or the MEF2 binding site, one finds, on average, one false positive every 120–130 bp [16,17]. Indeed, a promoter is characterized by an appropriate arrangement of TEs that allows specific interactions between the transcription factors that bind to them. During the past two years, four programs have been published that improve both the sensitivity and the specificity of promoter predictions by taking into account this complexity: the GRAIL-associated promoter prediction program [18•]; FunSiteP [19••]; PROMOTER SCAN [20••]; and PromFind [21••]. The GRAIL-associated promoter prediction program is in-

tended to recognize TATA box containing promoters; it combines statistical matrix scores and distance constraints for the TATA box, GC box, CAAT box, and cap sites using a neural network. An original feature of this system is the usage of protein coding region predictions provided by GRAIL to eliminate false promoter candidates. This program correctly predicts 66% of TATA box containing promoters, with one false positive every 23 kb [18•]. The interest of this program is limited, however, because many Pol II promoters do not contain the TATA box [20••]. Promoter predictions of FunSiteP and PROMOTER SCAN rely on the uneven distribution of TEs in promoter and nonpromoter sequences: regions of DNA that contain a higher density of putative TEs relative to nonpromoter sequences are more likely to be true promoter regions. PROMOTER SCAN uses the ratio of occurrence frequencies of a particular signal in promoter and nonpromoter regions as the primary input to the prediction algorithm, whereas FunSiteP takes an element's positional distribution relative to the transcription initiation site into account. Both methods correctly predict 60–70% of Pol II promoters, with about one false positive every 10 kb [19••,20••]. A possible drawback of these methods is that they rely upon extensive databases of known TEs. Hence, they may be unable to identify new promoter classes containing TEs that have not been yet characterized. PromFind also relies on the uneven distribution of TEs, but, instead of focusing on putative TEs, it analyzes the frequency of all possible hexamers in promoter and nonpromoter training sets. The specificity and sensitivity of PromFind and PROMOTER SCAN have been compared: both methods correctly predict 50–60% of vertebrate Pol II promoters, with about one false positive every 19 kb [21••].

Composite control elements

Despite significant improvements during the past two years, the success rate of promoter prediction programs is still relatively weak: 30–40% of true promoters are missed, and about 45–60% of predicted promoters are false positives. Future improvements will require better models of transcription-control regions that take into account not only the density but also the correct combination and spatial organization of TEs, and their position relative to other gene features. Kel *et al.* [22•] have compiled a database of composite elements (COMPEL) affecting gene transcription in vertebrates. On the basis of this information resource, an algorithm has been developed to locate potential composite elements in functionally uncharacterized DNA sequences [23••]. Quandt *et al.* [24•] developed a software package called GenomeInspector to detect potentially synergistic signals in genomes. Their system is able to assess distance correlations between many types of experimentally determined or predicted sequence features, for example, TEs, open reading frames, repeated elements, etc. The underlying assumption is that positionally correlated sequence elements are more probably associated with biological function than individual elements. Such a combination of biological knowledge with computer sequence analysis will allow the description of more accurate models of regulatory regions, as illustrated by recent examples on lentivirus LTRs [25] or muscle-specific promoters [26•].

Comparative sequence analysis: a powerful approach for the identification of new unknown regulatory elements

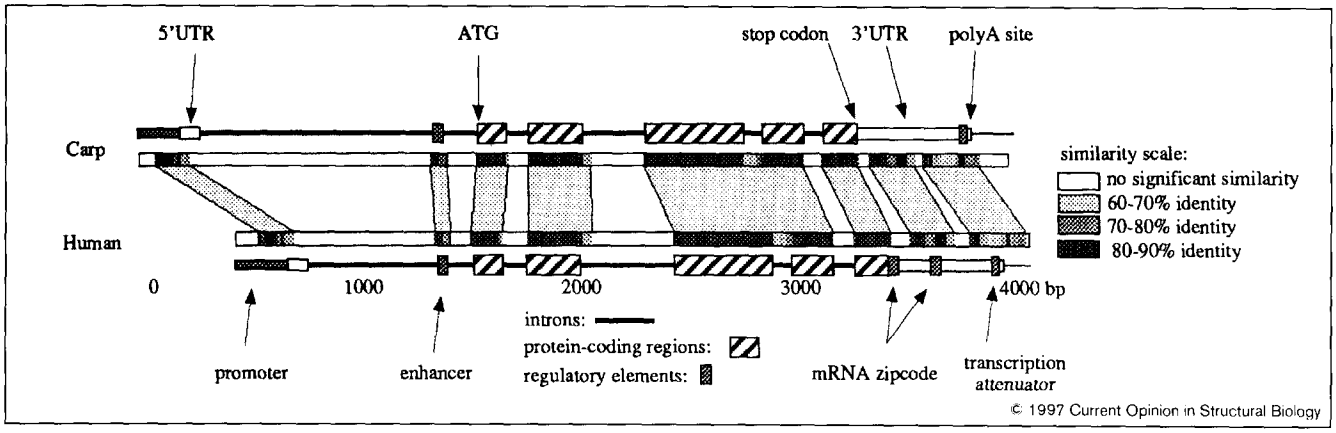
For most of the regulatory elements mentioned in the introduction, current biological knowledge is still insufficient to allow the development of prediction methods such as those described above for Pol II promoters. Moreover, it is probable that many regulatory elements are still totally unknown and remain to be discovered. In our opinion, the most promising approach for the identification of new unknown regulatory elements lies in the phylogenetic comparison of homologous sequences. This approach is not new, but it has recently gained considerable interest thanks to the start of projects intended to sequence large regions of some model vertebrate genomes [27,28••–30••].

Phylogenetic footprinting

The pattern of mutations that have occurred during evolution is an excellent indicator of functional constraints. Genomes continually undergo mutations, but the outcome of each mutation depends on its phenotypic effect. Mutations that are deleterious are generally eliminated by natural selection, whereas mutations that have no phenotypic effect (neutral mutations) or that are only slightly deleterious can be randomly fixed in the population (genetic drift). The consequence of this is that mutations accumulate much faster at nonfunctional DNA bases than at functionally constrained base positions. Hence, if one detects a sequence that has remained highly conserved during evolution, then it probably means that this sequence is functional (but the reverse proposal is not true: a sequence can be functional albeit nonconserved). Tagle *et al.* [31] proposed the term 'phylogenetic footprinting' to describe the phylogenetic comparisons that reveal evolutionary conserved functional elements in homologous genes. The efficiency of phylogenetic footprinting is illustrated in Figure 1, which shows the comparison of human and carp β-actin genes. This comparison shows that after >900 million years (Myrs) of divergence (450 Myrs in each lineage), four discrete elements in noncoding regions still remain highly conserved. Indeed, these four conserved noncoding regions correspond to essential regulatory elements that are involved in transcription and post-transcriptional processes (Fig. 1). Thus, the simple comparison of homologous sequences can reveal essential functional elements.

One should note, however, that regulatory elements that have been acquired very recently in evolution may not be detectable by phylogenetic footprinting. Moreover, the conserved feature does not necessarily reside in the primary sequence itself. In some cases, it is the spatial

Figure 1



Phylogenetic comparison of human and carp (*Cyprinus carpio*) β-actin genes. Sequences (accession numbers M24113 and M10277, respectively) have been compared using the LFASTA local similarity search program [44]. Regions of sequence similarity between the two genes are represented along with some sequence features (exons, introns, regulatory elements). As expected for a gene coding for a highly constrained protein, protein-coding regions are highly conserved. More surprising, after >900 Myrs of divergence (450 Myrs in each lineage), four discrete elements have remained highly conserved in noncoding regions. Indeed, these conserved noncoding regions have been shown to correspond to important regulatory elements: promoter elements in the 5'-flank [45]; an enhancer in the first intron [45]; a transcription attenuator in the 3'-end of the gene [46]; and 3'UTR elements involved in β-actin mRNA subcellular localization (mRNA zipcode) [47].

structure or a particular compositional property of the DNA or RNA that may be subject to selective pressure.

Choice of species for phylogenetic footprinting

The choice of species to be compared is essential for the efficiency of phylogenetic footprinting. If species are too closely related, distinguishing highly constrained regulatory elements from nonfunctional regions is impossible because there will not have been enough evolutionary time for the accumulation of mutations at neutral base positions. But if species are too distantly related, then detecting conserved regulatory elements may be impossible, either because they will have diverged too much to preserve any significant similarity or because the regulation processes are different in the two lineages.

Since the rate of accumulation of substitutions at neutral base positions has been estimated to be around 0.5% every Myrs [32], the sequence similarity between species that diverged 300 Myrs ago in DNA regions that are not subject to selective pressure should be about 30% (after correction for multiple substitutions), which is approximately the same as between two unrelated sequences. Any significant sequence conservation between species that diverged 300 Myrs ago should, therefore, indicate a strong selective pressure, and hence an important functional element. Many regulatory elements, however, have not been conserved for such a long time. The solution to detecting more recent regulatory elements consists in comparing more than two orthologous sequences. Species should be selected so that the cumulative length of branches of the phylogenetic tree uniting them to their last common ancestor represent >200 Myrs [33••]. The best picture can be obtained by comparing several species covering a wide range of evolutionary distances. This allows one to focus first on the most conserved elements that probably reflect essential regulatory processes shared by all species, and to progressively identify less conserved elements that may be involved in lineage-specific regulations.

Recent applications of phylogenetic footprinting

Since their original publication describing the phylogenetic-footprinting approach [31], Goodman's group has been using this method to identify elements involved in the developmental regulation of the mammalian β-globin cluster. After aligning orthologous sequences from several mammalian species, they search for sequence motifs with 100% conservation over at least six contiguous bps. These phylogenetic footprints are then analyzed by gel mobility shift essays to test whether they bind proteins. The efficiency of this approach has been first demonstrated on the γ-globin gene: of the 13 phylogenetic footprints identified, 12 (92%) correspond to binding sites of nuclear proteins, whereas only two out of nine nonconserved regions bind proteins [34]. Ultimately, 35 phylogenetic footprints have been detected in γ and ε genes, and their binding proteins have been identified [33••]. Other recent results of the phylogenetic-footprinting approach include the identification of promoter or enhancer elements in COX5B from primates [35], in Hoxb-1 [36••] and Hoxb-4 [37••] from vertebrates, and in IL-2Ralpha [38] from mammals.

Differential phylogenetic footprinting and motif-based phylogenetic footprinting

Two variants of the phylogenetic-footprinting method have been published. The first one, called 'differential phylogenetic footprinting', relies on a search for sequence

differences. This approach may be used to identify regulatory elements responsible for the establishment of novel expression patterns in specific lineages (see [33••], and references therein). The analysis is carried out by aligning orthologous sequences and searching for sequence differences. Probes spanning the sequence differences are then tested by gel mobility shift to detect differences in the pattern of proteins binding to them. This strategy was used to identify elements responsible for the specific expression pattern of γ-globin in primates (γ-globin is expressed in fetus in simian primates, whereas it is expressed in embryo in other mammals) [33••].

The second variant, called 'motif-based phylogenetic footprinting', has been developed to detect conserved binding sites that show sequence variation (see [33••], and references therein). Rather than focusing solely on primary sequence conservation, this method searches for putative TEs that occur at orthologous positions, allowing the detection of functionally conserved binding sites despite sequence differences.

Large-scale phylogenetic footprinting

While the examples described above focus on a few genes, the phylogenetic-footprinting approach can also be used for large-scale genome sequencing projects. In a first attempt, Duret *et al.* [39] have systematically analyzed noncoding sequences available in databases, in order to search for evolutionary conserved regulatory elements in vertebrate genes. More recently, we have extended this systematic comparative analysis to 145 Mb of noncoding sequences from different metazoan taxa (essentially vertebrates, insects and nematodes; L Duret, unpublished data). These large-scale phylogenetic footprinting analyses revealed hundreds of long noncoding elements that have remained highly conserved for 310–540 Myrs.

Three important results came out of these analyses. First, phylogenetic footprints are more frequent than could have been expected: at least 36% of orthologous genes between mammals and birds contain such highly conserved regions (HCRs) in their noncoding sequences after 310 Myrs of evolution (Table 1). However, the frequency of HCRs decreases as evolutionary distances increase: HCRs are two times more frequent between mammals and birds than between mammals and bony fishes (Table 1). The oldest HCR that has been observed predates the divergence between chordates and echinoderms (about 540 Myrs) and corresponds to the histones' 3′-processing signal. Despite a large amount of sequence data, we have not detected any significant conservation between vertebrates and insects or nematodes (L Duret, unpublished data). Thus, *Drosophila* and *Caenorhabditis elegans* are not suitable for finding phylogenetic footprints in the human genome.

Second, HCRs are almost two times more frequent in 3′-noncoding regions than in 5′-noncoding regions, and three times more frequent than in introns (Table 1). This observation is surprising because one would have expected a stronger selective pressure on 5′-regions of genes, as these are known to contain elements involved in regulation of transcription. Indeed, the analysis showed that most of these 3′HCRs are probably involved in post-transcriptional processes that are now recognized to be essential for regulating the expression of many genes [40]. Some of the detected HCRs correspond to already identified regulatory elements, but most of them are totally unknown.

Finally, the most surprising result is size of these HCRS: these conserved elements (at least 70% identity between species that diverged >300 Myrs ago) cover on average more than 400 nt (50–20 000 nt). None of the regulatory processes known to date can explain such a conservation

Table 1

Frequency of occurrence of evolutionary conserved elements in noncoding regions of other vertebrate genes*.

Gene region	Mammals / Birds 300 Myrs†	Mammals / Amphibians 350 Myrs†	Mammals / Bony fishes 450 Myrs†
3′-noncoding regions			
Number of orthologous genes‡	284	191	72
Frequency of genes containing conserved elements#	35.9% (102)	26.7% (51)	16.7% (12)
5′-noncoding regions			
Number of orthologous genes‡	96	51	25
Frequency of genes containing conserved elements#	19.8% (19)	3.9% (2)	16.0% (4)
Introns			
Number of orthologous genes‡	63	8	12
Frequency of genes containing conserved elements#	11.1% (7)	0% (0)	8.3% (1)

*To estimate the fraction of genes that contain evolutionary conserved elements, we searched for phylogenetic footprints within all orthologous genes between mammals and other vertebrate classes (selected from the HOVERGEN database [48]). †Approximate divergence time. ‡We have included only orthologous genes, for which at least 200 nt of noncoding region is available. #The minimal threshold to report conserved elements is 70% identity over 50 bp (L Duret, unpublished data).

over such a length. Most of regulatory elements that have been identified to date correspond to binding sites of regulatory proteins, and their sizes range from 5–25 nt. Do these long HCRs correspond to multiple adjacent binding sites? Or are they involved in more complex structure (e.g. at the RNA level)? The precise mechanisms in which these long HCRs are involved remain to be determined.

Conclusion

Methods of predicting Pol II promoters have been significantly improved during the past few years. Even if there is still room for improvements, one can note that, for the first time, the error rate of these methods is low enough such that the programs are of practical interest. Can the methodology developed for Pol II promoter prediction be transposed to other regulatory elements? Indeed, this approach is limited to regulatory elements that are relatively well characterized. The construction of rules for the prediction of translation start sites [41] or polyadenylation sites [18•] has been possible from the analysis of numerous experimentally defined examples. For most of the regulatory elements mentioned in the introduction, however, current biological knowledge is still insufficient to allow the development of any rule-based prediction method.

As illustrated in this review, the phylogenetic-footprinting approach is very efficient for identifying new regulatory elements, even those of totally unknown type, and even those that occur where one would have not expected to find them. This comparative approach considerably increases the amount of information that can be extracted from any genomic sequence. Thus, the sequencing of genomes of model vertebrates should be extremely valuable to the understanding of our own genome.

Koop and coworkers [27,28••,29••] have been promoting this 'comparative genomics' strategy by sequencing large genomic regions (20–100 kb) of orthologous loci in man and mouse. One of the interests of the mouse is that, as a mammal, it represents a model quite similar to humans in term of gene-expression patterns. 80 Myrs of divergence, however, may not be enough evolutionary time for the stringent detection of conserved regulatory elements: some large genomic regions evolve faster than others [28••,29••], and whether sequence conservation reflects selective pressure (i.e. functional constraints) or low mutability (e.g. lower sensibility to mutagens for some chromatin domains or better efficiency of DNA repair or proofreading) is not always clear.

The Japanese pufferfish (*Fugu rubripes*) is a good model vertebrate for comparative genomics because its genome is small (about 7.5 times smaller than the human genome — ≈ 400 Mb versus ≈ 3000 Mb), and yet it contains roughly the same set of genes as other vertebrates [30••]. Moreover, the evolutionary distance that separates mammals and bony fishes (450 Myrs) guarantees that only essential functional elements have remained conserved. Indeed, phylogenetic-footprinting analyses of Hox genes have demonstrated the usefulness of pufferfish for the identification of regulatory elements [36••,37••]. Many regulatory elements, however, have not remained conserved for such a long time: only 16% of orthologous genes between mammals and bony fishes contain conserved elements in their noncoding regions (Table 1).

In our opinion, the chicken genome represents the best compromise for phylogenetic footprinting: it has enough evolutionary distance (300 Myrs) but still many conserved elements (Table 1). In addition, it has a relatively small genome (1200 Mb) [42], and it has been suggested that 30–50% of its genes are concentrated in minichromosomes in which the gene density (~ 1 gene every 10 kb) approaches that seen in pufferfish [43]. Hence, sequencing of chicken minichromosomes could be of relatively low cost and yet could provide very valuable data for the identification of conserved regulatory elements in the human genome.

Acknowledgements

We thank Marc Robinson for a critical reading of the manuscript. When preparing this article, L Duret was recipient of a FEBS long-term fellowship.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fickett JW: **Finding genes by computer: the state of the art.** *Trends Genet* 1996, **12**:316–320.

2. • Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24**:238–241.
TRANSFAC includes data on eukaryotic transcription-regulating DNA-sequence elements and on the transcription factors binding to and acting through them. Currently, it is the only such database that is regularly updated.

3. Ghosh D: **Status of the transcription factors database (TFD).** *Nucleic Acids Res* 1993, **21**:3117–3118.

4. • Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol* 1996, **6**:361–365.
A concise and very useful introduction to hidden Markov model applications in molecular sequence analysis. A biological, not a mathematical paper.

5. Waterman MS, Arratia R, Galas DJ: **Pattern recognition in several sequences: consensus and alignment.** *Bull Math Biol* 1984, **46**:515–527.

6. Larsen NI, Engelbrecht J, Brunak S: **Analysis of eukaryotic promoter sequences reveals a systematically occurring CT-signal.** *Nucleic Acids Res* 1995, **23**:1223–1230.

7. • Pedersen AG, Baldi P, Brunak S, Chauvin Y: **Characterization of prokaryotic and eukaryotic promoters using hidden Markov models.** *Intel Sys Mol Biol* 1996, **4**:182–191.
A methodologically interesting paper. The results on eukaryotic promoters are difficult to evaluate.

8. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262**:208–214.

9. Fraenkel YM, Mandel Y, Friedberg D, Margalit H: **Identification of common motifs in unaligned DNA sequences: application**

to *Escherichia coli* Lrp regulon. *Comput Appl Biosci* 1995, 11:379–387.

10. • Wolfertstetter F, Frech K, Herrmann G, Werner T: **Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm.** *Comput Appl Biosci* 1996, **12**:71–80.
This paper describes a motif search algorithm for sets of functionally related DNA sequences using novel criteria to distinguish biologically relevant motifs from mismatched oligonucleotides over-represented by chance.

11. • Ulyanov AV, Stormo GD: **Multi-alphabet consensus algorithm for identification of low specificity protein–DNA interactions.** *Nucleic Acids Res* 1995, **23**:1434–1440.
A new algorithm is described that is capable of simultaneously extracting two motifs from a mixture of DNA sequence fragments binding to two different proteins.

12. Queen C, Wegman MN, Korn LJ: **Improvements to a program for DNA analysis: a procedure to find homologies among many sequences.** *Nucleic Acids Res* 1982, **10**:449–456.

13. • 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* 1995, **23**:4878–4884.
See annotation [14•].

14. • Chen QK, Hertz GZ, Stormo GD: **MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices.** *Comput Appl Biosci* 1995, **11**:563–566.
These two papers [13•,14•] present software tools to search for transcription factor binding sites using libraries of weight matrices. These libraries are constructed using data from TRANSFAC and TFD. The use of weight matrices instead of IUPAC code based consensus sequences significantly improves the accuracy of predictions.

15. • Prestridge DS: **SIGNAL SCAN 4.0—additional databases and sequence formats.** *Comput Appl Biosci* 1996, **12**:157–160.
This software tool searches for transcription factor binding sites using TFD and TRANSFAC. In addition to using IUPAC code based consensus sequences, the current version now includes the MATRIX SEARCH software described above [14•].

16. Prestridge DS, Burks C: **The density of transcriptional elements in promoter and non-promoter sequences.** *Hum Mol Genet* 1993, **2**:1449–1453.

17. Fickett JW: **Quantitative discrimination of MEF2 sites.** *Mol Cell Biol* 1996, **16**:437–441.

18. • Matis S, Xu Y, Shah M, Guan X, Einstein JR, Mural R, Uberbacher E: **Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence.** *Comput Chem* 1996, **20**:135–140.
See annotation [21••].

19. •• Kondrakhin YV, Kel AE, Kolchanov NA, Romaschenko AG, Milanesi L: **Eukaryotic promoter recognition by binding sites for transcription factors.** *Comput Appl Biosci* 1995, **11**:477–488.
See annotation [21••].

20. •• Prestridge DS: **Predicting Pol II promoter sequences using transcription factor binding sites.** *J Mol Biol* 1995, **249**:923–932.
See annotation [21••].

21. •• Hutchinson GB: **The prediction of vertebrate promoter regions using differential hexamer frequency analysis.** *Comput Appl Biosci* 1996, **12**:391–398.
These four papers [18•,19••–21••] describe algorithms and software tools for the prediction of Pol II promoters. For the first time, the error rate is low enough to make them of practical interest.

22. • Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA: **A compilation of composite regulatory elements affecting gene transcription in vertebrates.** *Nucleic Acids Res* 1995, **23**:4097–4103.
This paper describes a potentially very useful database of experimentally characterized synergistic and antagonistic genetic element pairs.

23. •• Kel AE, Kondrakhin YV, Kolpakov PHA, Kel OV, Romaschenko AG, Wingender E, Milanesi L, Kolchanov NA: **Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences.** *Ismb* 1995, **3**:197–205.
Several recently developed algorithms are described that analyse gene regulatory regions including prediction methods for promoters and composite elements.

24. • Quandt K, Grote K, Werner T: **GenomeInspector: basic software tools for analysis of spatial correlations between genomic structures within megabase sequences.** *Genomics* 1996, **33**:301–304.
This program is able to assess distance correlations between many types of experimentally determined or predicted sequence features. It is an useful tool to search for potentially synergistic signals and thus construct models of regulatory regions.

25. Frech K, Brack-Werner R, Werner T: **Common modular structure of lentivirus LTRs.** *Virology* 1996, **224**:256–267.

26. • Fickett JW: **Coordinate positioning of MEF2 and myogenin binding sites.** *Gene* 1996, **172**:GC19–GC32.
An interesting case study is presented on the naturally occurring spacing between two classes of transcription regulatory elements recognized by two synergistically acting DNA-binding proteins.

27. Hood L, Koop B, Goverman J, Hunkapiller T: **Model genomes: the benefits of analysing homologous human and mouse sequences.** *Trends Biotechol* 1992, **10**:19–22.

28. •• Koop BF: **Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution.** *Trends Genet* 1995, 11:367–371.
See annotation [29••].

29. •• Koop BF, Richards JE, Durfee TD, Bansberg J, Wells J, Gilliam AC, Chen HL, Clausell A, Tucker PW, Blattner FR: **Analysis and comparison of the mouse and human immunoglobulin heavy chain J_H-Cm-Cd locus.** *Mol Phylogenet Evol* 1996, **5**:33–49.
These two papers [28••,29••] show that large portions of mammalian genomes evolve at different rates: some loci are highly conserved; and some quickly diverge, whereas others show a mixed pattern. Whether these different patterns reflect variations in selective pressure is unclear. Rather, the authors suggest that the genomic context may affect the local rate of evolution. An important conclusion is that sequence conservation between man and mouse does not necessarily imply a function.

30. •• Elgar G, Sandford R, Aparicio S, MacRae A, Venkatesh B, Brenner S: **Small is beautiful: comparative genomics with the pufferfish (*Fugu rubripes*).** *Trends Genet* 1996, **12**:145–150.
A good review on the features of the *Fugu* genome and its usefulness for comparative analysis.

31. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT: **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* 1988, **203**:439–455.

32. Li WH, Luo C, Wu C: **Evolution of DNA sequences.** In *Molecular Evolutionary Genetics.* Edited by Macintyre RJ. New York: Plenum Press; 1985:1–94.

33. •• Gumucio DL, Shelton DA, Zhu W, Millinoff D, Gray T, Bock JH, Slightom JL, Goodman M: **Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the β-like globin genes.** *Mol Phylogenet Evol* 1996, **5**:18–32.
The authors present the principles of phylogenetic footprinting, differential phylogenetic footprinting and motif-based phylogenetic footprinting, and they show the efficiency of these approaches in identifying regulatory elements in mammalian β-globin clusters.

34. Gumucio DL, Blanchar-Mcquate K, Heillstedt-Williamson H, Tagle DA, Gray TA, Tarle SA, Gragowski L, Goodman M, Slightom JL, Collins FS: **γ-globin gene regulation: evolutionary approaches.** In *The Regulation of Hemoglobin Switching, Proceedings of the Seventh Conference on Hemoglobin Switching.* Edited by Stamatoyannopoulos G, Nienhuis AW. Baltimore: John Hopkins University Press; 1991:277–289.

35. Bachman NJ, Yang TL, Dasen JS, Ernst RE, Lomax MI: **Phylogenetic footprinting of the human cytochrome *c* oxidase subunit Vb promoter.** *Arch Biochem Biophys* 1996, **333**:152–162.

36. •• Popperl H, Bienz M, Studer M, Chan SK, Aparicio S, Brenner S, Mann RS, Krumlauf R: **Segmental expression of Hoxb-1 is controlled by a highly conserved autoregulatory loop dependent upon exd/pbx.** *Cell* 1995, **81**:1031–1042.
See annotation [37••].

37. •• Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S: **Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*.** *Proc Natl Acad Sci USA* 1995, **92**:1684–1688.
These two papers [36••,37••] illustrate the usefulness of *Fugu* for the identification of conserved regulatory elements. Conserved elements detected by comparative sequence analysis are shown to act as enhancers using deletion analyses in the murine loci, as well as directly testing the *Fugu* sequences using transgenesis in mice.

38. Lecine P, Algarte M, Rameil P, Beadling C, Bucher P, Nabholz M, Imbert J: **Elf-1 and Stat5 bind to a critical element in a new enhancer of the human interleukin-2 receptor alpha gene.** *Mol Cell Biol* 1996, **16**:6829–6840.

39. Duret L, Dorkeld F, Gautier C: **Strong conservation of non-coding sequences during vertebrates evolution – potential involvement in post-transcriptional regulation of gene expression.** *Nucleic Acids Res* 1993, **21**:2315–2322.

40. Jackson RJ: **Cytoplasmic regulation of messenger RNA function – the importance of the 3′ untranslated region.** *Cell* 1993, **74**:9–14.

41. Kozak M: **Interpreting cDNA sequences – some insights from studies on translation.** *Mamm Genome* 1996, **7**:563–574.

42. Tiersch TR, Wachtel SS: **On the evolution of the genome size of birds.** *J Heredity* 1991, **82**:363–368.

43. MacQueen HA, Fantes J, Cross SH, Clark VH, Archibald AL, Bird AP: **CpG islands of chicken are concentrated on minichromosomes.** *Nat Genet* 1996, **12**:321–324.

44. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85**:2444–2448.

45. Liu ZJ, Moav B, Faras AJ, Guise KS, Kapuscinski AR, Hackett PB: **Functional analysis of elements affecting expression of the beta-actin gene of carp.** *Mol Cell Biol* 1990, **10**:3432–3440.

46. Deponti-Zilli L, Seiler-Tuyns A, Paterson BM: **A 40-base-pair sequence in the 3′ end of the β-actin gene regulates β-actin mRNA transcription during myogenesis.** *Proc Natl Acad Sci USA* 1988, **85**:1389–1393.

47. Kislauskis EH, Zhu XC, Singer RH: **Sequences responsible for intracellular localization of beta-actin messenger RNA also affect cell phenotype.** *J Cell Biol* 1994, **127**:441–451.

48. Duret L, Mouchiroud D, Gouy M: **HOVERGEN: a database of homologous vertebrate genes.** *Nucleic Acids Res* 1994, **22**:2360–2365.