

Expert Opinion

1. Introduction
2. Regulation of drug-metabolising enzymes by transcription factors
3. Approaches to study transcriptional regulation
4. Conclusion
5. Expert opinion

For reprint orders,
please contact:
ben.fisher@informa.com

informa
healthcare

Prediction of *cis*-regulatory elements for drug-activated transcription factors in the regulation of drug-metabolising enzymes and drug transporters

Michael Podvinec[†] & Urs A Meyer

[†]*Swiss Institute of Bioinformatics and Biozentrum, University of Basel, Klingelbergstrasse 50-70, CH-4056 Basel, Switzerland*

The expression of drug-metabolising enzymes is affected by many endogenous and exogenous factors, including sex, age, diet and exposure to xenobiotics and drugs. To understand fully how the organism metabolises a drug, these alterations in gene expression must be taken into account. The central process, the definition of likely regulatory elements in the genes coding for enzymes and transporters involved in drug disposition, can be vastly accelerated using existing and emerging bioinformatics methods to unravel the regulatory networks causing drug-mediated induction of genes. Here, various approaches to predict transcription factor interactions with regulatory DNA elements are reviewed.

Keywords: drug induction, drug-metabolising enzymes, drug response elements, drug transporters, nuclear receptors, transcription factors, transcription factor binding

Expert Opin. Drug Metab. Toxicol. (2006) 2(3):367-379

1. Introduction

A large number of drugs and other xenobiotics are known to affect the expression of drug-metabolising enzymes and transporters (for recent reviews refer to [1-3]). In doing so, they may alter their own metabolism, as well as that of other compounds cleared by the same enzyme or transporter system. Transcription factors, in particular of the nuclear receptor and the basic helix-loop-helix families, mediate these changes of gene expression. This review discusses bioinformatics approaches to investigate genomic sequences, and how binding sites of transcription factors likely to be involved in the regulation of target genes can be identified.

Over the last decade, biology has entered an age where genomic information is commonplace. At present, the complete sequence of 373 genomes is available, among which 41 are from eukaryotes, including many organisms of clinical or research interest [101]. In addition to these freely available resources, researchers may have access to further proprietary or licensed sequence databases. Today, the challenge, therefore, is how to efficiently extract meaning from this host of information.

In the press conference that announced the results of the human genome sequencing effort, Bill Clinton proclaimed, as a tribute to Galileo, 'Today, we are learning the language in which God created life' [102]. The picture that this statement evokes strikingly reflects many of the problems faced with the deciphering of genomic information. Indeed, looking at the raw sequence of nucleotides making up a genome is like looking at an unknown language, for which we know neither the alphabet nor the words, and even less what they may mean.

The process of understanding the information can be characterised by the following steps of a cycle. First is the identification of elements; in order to extract meaning from

a sequence, we need to annotate, structure and classify the information contained within. Similar to how we might decipher a foreign text by discerning certain letter combinations and tentatively assigning words, known and inferred information is combined with genome sequences in the form of annotation of known or putative genes and exons, allelic variants, and so on. All these pieces of primary information are stored in large databases, and are made available through interfaces such as Ensembl [4,103] or the UCSC Genome Browser [5,104], and can be used as stepping stones in inferring further information from the genome. Second is assigning function to elements; the next level of understanding genes is the assignment of particular functions in the cellular context, similar to the translation of words in an unknown text. Third is unravelling the context; ultimately, contextual information, in the form of functional, metabolic and regulatory networks between different genes and proteins, needs to be interpreted in order to understand not only the individual function of a gene, but its function within the context of the system. This is the challenge that systems biology is beginning to address [6,7]. Understanding the regulatory and metabolic context is the next necessary step to further our understanding of the information produced in genome-sequencing projects.

This paper focuses on the regulatory aspect of gene context, specifically on the transcriptional regulation of drug-metabolising and -transporting enzymes. In particular, bioinformatics methods are described that are available to assist in the study of these regulatory processes.

2. Regulation of drug-metabolising enzymes by transcription factors

At its most fundamental, transcriptional control is exerted by means of transcription factors binding to target sites within a gene (e.g., in the untranslated regions or in introns) or around a gene (upstream or downstream). In response to endogenous or exogenous signals, transcription factors become activated and bind to their DNA response element through specific DNA-binding domains, such as the basic helix-loop-helix/PAS motif found in the aryl hydrocarbon receptor [8] or a three-helix zinc-finger motif found in nuclear receptors [9]. Nuclear receptors such as the pregnane X receptor (PXR) and the constitutive androstane receptor (CAR) are key transcription factors that mediate the induction of clinically relevant CYPs, in particular of the CYP2, -3 and -4 families, as well as of drug transporters (reviewed in [3,10]). Following activation or nuclear translocation by drugs, other xenobiotics and endogenous compounds, PXR and CAR form heterodimers with the retinoid X receptor (RXR) and bind to the *cis*-regulatory elements described below. Their binding to these DNA sequences results in the recruitment of co-activator proteins, accessory proteins that cause chromatin decondensation and thus render the DNA more accessible to other transcription factors and the transcription initiation complex. Moreover, members of the activation complex signal

directly to the transcription machinery, causing enhanced transcription of the gene at the transcription start site. There is extensive crosstalk between PXR, CAR and other nuclear receptors of the same family, namely with bile acid receptor FXR (farnesoid X receptor), the sterol receptor liver X receptor (LXR) and the vitamin D receptor [11].

To predict whether a particular transcription factor regulates a certain gene, it is thus necessary to investigate the presence of response elements within the gene. This is a difficult task, as transcriptional activation of a target gene cannot simply be modelled by activation or suppression through transcription factors, but rather it depends on a multitude of factors, such as competitive or cooperative binding of proteins to DNA elements [12], chromatin bending and further, context-dependent interactions (reviewed in [13]), which are not directly tractable at present.

Moreover, prediction of transcription factor-binding sites is hindered by the fact that the core sequence elements recognised by transcription factors are short and exhibit considerable sequence variability. Composed of only four different nucleic acids, DNA sequences have lower information content compared with peptide sequences. This makes motif discovery and detection much harder than in proteins, and any algorithm to find such ill-defined, short sites will invariably return numerous false positive results.

To discern regulatory elements within DNA sequences specifically, additional features of these elements must be considered. Most importantly, transcription factor-binding sites have been observed to cluster together in *cis*-regulatory modules [13,14]. These comprise response elements of several transcription factors, and serve as information-processing devices, effectively integrating multiple inputs and cross-connecting signal transduction pathways [15]. Such regulatory modules are 300 – 500 bp long and often conserved in the same gene in different species, as well as between genes involved in the same pathways or responses. In addition, noncoding regions of the genome containing *cis*-regulatory modules show higher interspecies conservation than neutral intergenic regions.

3. Approaches to study transcriptional regulation

A popular way to investigate which transcription factors influence the expression of a given gene, or, inversely, which genes may be influenced by a particular transcription factor (or a set) is to use a computer program to look for response elements of the transcription factor in question. Various methods and strategies for this search, as well as various postfiltering methods, have been published and for the most part are available to use as through a web portal or as a standalone program. Often, however, it is left to the user how to best apply prior knowledge in choosing a combined approach that attains desired specificity and sensitivity levels.

Table 1. IUB-IUPAC DNA ambiguity codes (used to express consensus sequences, where one position may stand for more than one kind of nucleotide).

| Ambiguity codes | Corresponding nucleotides |
|-----------------|---------------------------|
| A | Adenine |
| C | Cytosine |
| G | Guanine |
| T | Thymine |
| M | A or C |
| R | A or G |
| W | A or T |
| S | C or G |
| Y | C or T |
| K | G or T |
| V | A or C or G |
| H | A or C or T |
| D | A or G or T |
| B | C or G or T |
| X/N | A, C, G or T |

The techniques described below lend themselves to combination with other bioinformatics methods, as well as with independent experimental results from, for example, expression array experiments or studies performed in other model systems. The more information is collected before embarking on a project, and the better the binding sites of interest are modelled and understood, the more meaningful the outcome will be. As a starting point, the binding site of a transcription factor must be represented in a way to be recognised efficiently in target sequences.

3.1 Modelling transcription factor-binding preference

Fundamentally, binding sites consist of short sequences, up to a length of 20 bp. Functional binding sites can be variable in sequence; at most positions of the binding site, base exchanges are possible without affecting the functionality of the binding site, as observed, for example, in [16] or [17]. At the same time, particular positions within the binding site, when altered, abolish or strongly reduce binding of the transcription factor. Transcriptional regulation is thus based on a very flexible system that recognises a variety of sequences, but even so, constraints are put on what constitutes a functional binding site.

A given transcription factor's preference for binding site sequences must be modelled in a way in which searching is efficient, in which the model is specific, but takes the variability of experimentally verified binding sites of this transcription factor into account, and finally a way in which the number of false detections is kept low. There are several common ways of achieving these goals, which are briefly discussed.

3.1.1 Patterns: consensus sequences and regular expressions

This way of representing the common denominator of a collection of aligned sequences is most familiar to any biologist. That in itself is a major advantage of consensus sequences – they are very accessible and easily understood. Bases within a consensus sequence are represented by IUB-IUPAC DNA ambiguity codes (Table 1).

Given the example sequences, AGTTCA, AGGTGT, AGTCTT and AGTCAA, we align the sequences and derive the consensus sequence as AGKYNW. Consensus sequences are an efficient way to represent information for searching in large sequence databases, as computer programs can deal with this representation very easily and quickly. On the other hand, consensus sequences are not well suited to represent DNA binding sites for proteins [18,19]. When a consensus sequence is created, information about the frequency of nucleotides at a given position is lost. For example, a position where both thymine and adenine nucleotides are observed will be marked with the 'W' symbol, regardless of whether the two nucleotides occur with the same frequency, or whether one of them was found in the overwhelming majority of cases. Moreover, querying sequences for binding sites using a consensus sequence can only lead to match/no match decisions; no quantitative score of how well a particular sequence fits the consensus can be derived, thus preventing any attempt to rank and prioritise matches.

Consensus sequences are fine-tuned by modifying the underlying rule set. For instance, a position where A was found nine times, and T only once would be better represented by A than by W. Likewise, positions with only a slight preference for a specific nucleotide are better represented by N. Moreover, it may be advisable to ignore rare variations, and thus to remove outliers from the consensus building step. Neither of these strategies, however, can solve the fundamental lack of a useful scoring function. Advantages, limitations and applications of consensus sequences are summarised in Table 2.

Consensus sequences are simple examples of regular expressions, a computer science concept. They allow the definition of variable patterns in texts or in sequences and follow a given set of semantic rules. One example for a more elaborate rule set is the prosite syntax used to describe patterns in proteins (see [105]), another example is the regular expression syntax used in programming languages, such as Perl (Table 3). Regular expressions combine an alphabet of characters (such as the letters of the alphabet, the nucleic acid symbols, or the one-letter amino acid code) with a number of grouping characters, special characters and characters that define logical operations and repetition. Regular expressions are more powerful than consensus sequences due to their notion of quantifiers and modifiers, allowing for gaps, insertions, variable spacing and variable sequence lengths. For instance, a pattern to recognise the mRNA polyA tail, defined as the polyadenylation signal plus a tail of up to 250 adenines

Table 2. Advantages, limitations and applications of binding site models.

Consensus sequences

| | |
|--------------|--|
| Advantages | Simplicity: consensus sequences are intuitive and easily created Well-known: good to communicate with other researchers |
| Limitations | Qualitative: binary match/no match decision, no differentiated scoring Lossy: creation of a consensus discards information Delicate: very dependent on the chosen training set No possibility to accommodate gaps |
| Applications | Useful whenever the sequence feature you are searching for is extraordinarily similar; for example, restriction sites |

Patterns/regular expressions

| | |
|--------------|--|
| Advantages | Straightforward: easy to create and easy to understand Flexible: patterns can handle insertions, deletions and variable repetition |
| Limitations | Qualitative: binary match/no match decision, no scoring Lossy: creation of a pattern discards information Effectivity: too many variable positions will quickly allow almost anything to match |
| Applications | Searching for reasonably similar, relatively short, sequence features Patterns are often used to describe features of proteins |

Positional weight matrices

| | |
|--------------|---|
| Advantages | Specificity: frequency information from the multiple sequence alignment is retained Quantitative: scoring function allows differentiated analysis, not just yes/no answers |
| Limitations | Indels: insertions and deletions cannot be incorporated into the model |
| Applications | Currently method of choice for variable DNA sequences, such as transcription factor-binding sites |

Generalised profiles and HMMs

| | |
|--------------|--|
| Advantages | Indels: can deal with insertions and deletions Penalties can be precisely defined Sensitive: able to detect weak homologies Scoring: good, robust scoring system Reliability: HMMs are solidly based on probability theory |
| Limitations | Software: requires sophisticated software Expertise: significant <i>a priori</i> knowledge needed to use competently and successfully. The theoretical foundation of HMMs is complex, and requires perseverance from the user |
| Applications | Detection of motifs in proteins and nucleic acids |

HMM: Hidden Markov Model.

attached 10 – 20 bases downstream, can be defined like this: /AAUAA(A|C|G|U){10,20}(A){,250}\$/. The binding site of a transcription factor dimer that binds to two copies of the AGKYNW consensus separated by four or five nucleotides is described by: /AG(G|T)(T|C)(A|C|G|T)(A|T)(A|C|G|T){4,5}AG(G|T)(T|C)(A|C|G|T)(A|T)/. Advantages, limitations and applications of regular expression pattern searches are summarised in Table 2.

3.1.2 Nucleotide distribution matrices

Nucleotide distribution matrices, position-specific scoring matrices, positional weight matrices (PWMs) and nucleotide weight matrices are closely related concepts. Matrix models

solve one of the problems inherent with pattern representations: specific information about the distribution of nucleotides in the sequence is retained, and for each position of the binding site we can look up how likely a particular nucleotide is to occur in the binding sites the matrix was modelled after. Moreover, matrix models provide a scoring function that allows for the comparison of matches that were found in a sequence. Using the example from above (four binding sites: AGTTCA, AGGTGT, AGTCTT and AGTCAA), a matrix would be created in the following way: after aligning the sequences, the number of times an A, C, G or T nucleotide is present is counted for every column. For each nucleotide, its frequency is then entered into the matrix (the light grey box

Table 3. Some special characters and delimiters in Perl-like regular expression syntax.

| Character | Meaning |
|-----------|---|
| ^ | Start of a sequence |
| . | Any character |
| \$ | End of a sequence |
| | Separates alternatives |
| () | Encloses a subpattern |
| ? | Preceding expression may be there, or not |
| + | Preceding expression must be there at least once, but can be repeated |
| * | Preceding expression can be absent, or present once or many times |
| {n} | Match preceding expression exactly n times |
| {n, } | Match preceding expression $\geq n$ times |
| { ,n} | Match preceding expression 0 – n times |
| {n,m} | Match preceding expression n up to m times |

Refer to the text for some examples of how queries can be constructed using these symbols in conjunction with normal characters. The list is incomplete, a full introduction into regular expressions in Perl is provided in [129].

| | | | | | | |
|------------|------|------|------|------|------|------|
| Sequence 1 | A | G | T | T | C | A |
| Sequence 2 | A | G | G | T | G | T |
| Sequence 3 | A | G | T | C | T | T |
| Sequence 4 | A | G | T | C | A | A |
| A | 1.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.50 |
| C | 0.00 | 0.00 | 0.00 | 0.50 | 0.25 | 0.00 |
| G | 0.00 | 1.00 | 0.25 | 0.00 | 0.25 | 0.00 |
| T | 0.00 | 0.00 | 0.75 | 0.50 | 0.25 | 0.50 |
| Weight | 100 | 100 | 59 | 50 | 0 | 50 |

Figure 1. Creation of a weight matrix from four response element sequences. From four exemplary hexameric sequences, a nucleotide distribution matrix containing the frequencies of each nucleotide at each position was calculated (light grey box). In addition, positional weights were calculated for each position within the matrix (grey box).

in Figure 1). Nucleotide distribution matrices created this way are for example used in the Nuclear Receptor Binding Site Scanner (NUBIScan), developed by the group of Podvinec *et al.* [20], or in the MatInspector [21] algorithm. Other algorithms use a log-odds score of the nucleotide frequencies. This is obtained by dividing the observed frequency by the expected (background) frequency of the particular base. Finally, the logarithm of the odds score is entered into the matrix (e.g., see [22]). As training sets are always limited in size, and do not reflect all positive binding sites, many approaches introduce pseudocounts to circumvent the problem where a base was never encountered at a particular position within the training set, but cannot be excluded to occur in functional binding sites. A simple pseudocount method works on the assumption that all bases occur at least once at each position, whereas more sophisticated methods take the background distribution of nucleotides into account.

Nucleotide distribution matrices can be further refined by introducing a positional weight. The rationale behind this approach is that not all positions within a binding site are equally important for binding. Whereas at some positions, the chemical identity of the nucleotide may be crucial, at others, there may not be direct interaction with the transcription factor, and the nucleotide may merely be important for spacing. As a measure of how important a particular position within the context of the whole binding site is, we can determine the sequence conservation of that position among a set of aligned binding sites. A position at which all four nucleotides are found to occur equally as likely in functional binding sites is expected to be less important for binding than a position at which only one or two particular nucleotides are found. Information theory provides the theoretical framework to determine the information content of a particular signal [23]. The degree of sequence conservation at each position of the matrix

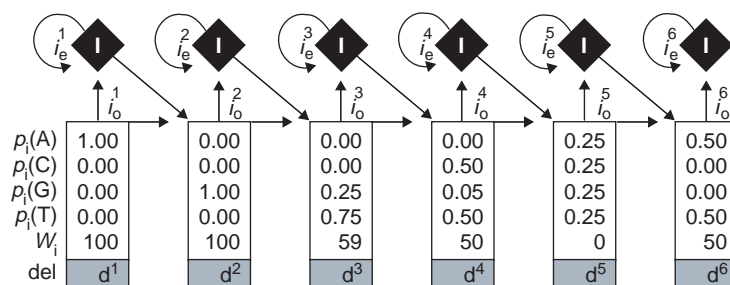


Figure 2. Extending the weight matrix to a generalised profile. In addition to the probability values p_i for each base at each position, and the weight W_i at that position, each position also has a score (d^1 to d^6) that is attributed when this position is deleted. Moreover, i_o is a score attributed for a gap opening and i_e , the first base of an insertion. Lastly, the score i_e is applied for any additional insertion. Scoring thus follows the arrows, taking either a score from the boxes, or one of the transition scores associated with insertions.

can be quantified through the information content of the nucleotide frequencies at that position. This value can be expressed in different ways; the weight values shown in **Figure 1** are calculated as described in [20], and are scaled in the range of 0–100; Equation 1:

(1)

$$W_i = \frac{100}{\ln 4} \left(\sum_{b=A,C,G,T} p_i(b) \ln p_i(b) + \ln 4 \right)$$

0, if $p_i(b) = 0$

Here, W_i signifies weight at position i of the matrix, b stands for any of the four bases A, C, G and T, and $p_i(b)$ is the frequency of the base b at position i . This equation is a variation of the one used in MatInspector [21]. Many other algorithms for matrix generation express positional weights as bits, with the maximum value being 2 bits for DNA sequences. These values are calculated as the difference between the maximum possible entropy and the entropy of the observed symbol distribution, according to Equation 2 [24]:

(2)

$$R_{seq} = S_{max} - S_{obs} = \log_2 4 - \left(- \sum_{b=A,C,G,T} p_b \log_2 p_b \right)$$

Here, 4 is the number of symbols in the DNA alphabet, and p_b is the observed frequency for each base at a particular sequence position.

Introducing positional weight is a common strategy to improve predictions in a number of matrix-based algorithms for the detection of transcription factor-binding sites. To search a sequence for hits to the matrix, all possible subsequences of the same length as the matrix are compared with the matrix by scanning the matrix along the query sequence.

To each subsequence, a score is attributed based on how frequent each nucleotide of that subsequence was observed at its corresponding position during matrix creation. The positional weight is taken into account, so that positions with less variation throughout the training set contribute more to the final score. Finally, hits that exceed a particular threshold are reported. Advantages, limitations and applications of nucleotide distribution matrices and positional weight matrices are summarised in **Table 2**.

Whereas weight matrices are designed to allow computers to perform efficient searches of query sequences, a related concept, the sequence logo, represents the same key information in a more visual and intuitive manner [24]. Here, the bases found in an aligned collection of binding sites are displayed from left to right as stacks of the letters A, C, G and T. The height of each letter is relative to its frequency at that position, and the overall height of the stack is relative to the information content of the particular position. A web server generating sequence logos is available [25,106].

3.1.3 Generalised profiles and hidden Markov models

Generalised profiles can be considered an extension of matrices that allows for insertions and deletions between bases. Using generalised profiles, DNA motifs can be modelled accurately starting from sets of true positive sequences [17]. At every position of the matrix, not only scores for each base are defined, but also scores for gap opening (the first base of an insertion), gap elongation (any further insertion), and a score if that position of the matrix is deleted (**Figure 2**). These insertion and deletion scores are usually penalty scores, that is, if insertion or deletion is detected, it will usually lessen the score. A common, publicly available suite of programs is pftools V2.0, written by Bucher [107]. It contains tools to create and calibrate profiles, and to perform searches with them.

Finally, generalised profiles can be transformed into hidden Markov models (HMMs) [26,27]. HMMs are a formal foundation for making probabilistic models of assigning features to parts of linear sequences. An HMM is composed of distinct nodes, or states that the system it describes can be in. It is

probably best envisaged not as a method for detecting features, but rather as a generator that can produce sequences like the one we are testing for. There are paths through the model that connect the individual nodes and that are associated with transition probabilities. In addition, each node has distinct emission probabilities – the likeliness of emitting a particular signal (in our case the bases A, C, G and T) when the process is in the specific state associated with the node. A path through the model changes the state it is in, for example, from 'no binding site' to 'first base of a likely binding site', from 'first base of a binding site' to 'second base of a binding site', and so forth. In each state associated with emission of a signal, signals will be generated according to the emission probabilities of that state. Both the transition probability and emission probability parameter set are assigned from training sequences. Using a dynamic programming algorithm, the most probable state path through the model given a query sequence and an HMM is then found. HMMs are full probabilistic models. This fact allows the application of Bayesian probability theory to analyse results obtained and to optimise the parameters of the model. One common toolset for doing analyses using HMMs is the HMMer suite [108]. Moreover, HMMs are used in a number of algorithms to detect transcription factor-binding sites (see below). Advantages, limitations and applications of generalised profiles and HMMs are summarised in Table 2.

3.2 Discovery of binding sites

By far the largest part of binding site discovery is still done by classical, wet-lab experimental methods, such as electrophoretic mobility shift assays (EMSA), site-directed mutagenesis and DNA-affinity purification, and not by initial computational screening. The functional characterisation of a binding site and of the transcription factor it interacts with, along with information about the physiological role of this transcription factor, is the gold standard for a binding site model.

Large databases are available that contain information about transcription factors taken from the scientific literature. In these, information about transcription factors and the sites recognised by them is then annotated and stored. The best-known example, and probably the most extensive of these, is the commercial TRANSFAC database [28], available from Biobase [109]. An older version is also freely accessible through Sequence Retrieval System (SRS) as the TFMATRIX databank [110]. Another database for matrix models, JASPAR, is freely available at [111] and focuses on nonredundancy and curation for improved quality [29]. Another freely available database is TRRD, which collects experimentally confirmed information about regulatory regions of genes, including transcription factor-binding sites [30,112]. Still, the number of sites catalogued in these databases is clearly just a small fraction of the total number of transcription factor-binding sites present in the genome.

The typical manner in which transcription factor-binding sites are discovered is serendipitous. Most of the time, binding sites are still discovered one by one, as the regulation of a particular gene is analysed. These data are then collected from

the literature, and used for binding-site predictions. From such collections of binding sites for a transcription factor, weight matrices or profiles are then created. Only relatively few studies have been done on mammalian transcription factors; in particular on the ones implied in the regulation of drug-metabolising enzymes, where the express purpose was the generation of transcription factor-binding site models. One such technique, SELEX (for SElective EXpansion) works by initially creating a random population of short, double-stranded DNA oligomers, followed by several cycles of selecting oligomers that bind to the transcription factor under study, for example, using EMSA experiments, and amplifying this selected population, so that good binders become enriched in each cycle. After a number of cycles, the resulting oligomers are sequenced and the sequences used to build a model for the binding site, such as in [17]. Such approaches can eliminate a shortcoming often encountered with binding-site models: true, verified binding sites are usually scarce, even after intensive literature study, and the training set of sequences is most likely small and in the statistical 'grey area'. Moreover, the true positive binding sites found in the literature are not obtained from an unbiased, random study, but rather were found in genes that were analysed in another context than expressly for model building. Care, however, is necessary to prevent overselection for pure strength of interaction, as binding strength, rather than functionality, is the selection criterion in SELEX.

3.2.1 Computational approaches to find novel binding sites

Several algorithms have been proposed that are able to pick up statistically over-represented patterns within sequences. MEME [31,113] is based on the expectation-maximisation algorithm, and returns nucleotide count matrices of found motifs. It can be used online or locally. The Gibbs Motif Sampler [32,114] is another tool, using Gibbs sampling to discover conserved elements between sequences. MDScan [33,115] applies a protocol of word enumeration from high-confidence sequences, followed by matrix updating using a whole set of sequences obtained from ChIP-chip or expression-array experiments. Teiresias [34,116] is a general-purpose pattern discovery algorithm used in fields ranging from pattern discovery in DNA or proteins to text mining and market analysis. In general, these unbiased methods require a set of sequences assumed to share at least one common transcription factor-binding site, even though many methods can cope with sequence sets where not every sequence has a particular site. Such a set of genes may be a collection of orthologues of the gene of interest from other species, or a set of genes shown to respond in the same fashion to a stimulus, for instance, in an expression-array experiment. These two criteria can be combined for more specificity, selecting for motifs found in orthologues of coregulated genes.

3.3 Knowledge-based detection of binding sites

In the majority of use cases, bioinformatics approaches are used to detect the presence of binding sites for any of a

number of known transcription factors within a gene, based on knowledge imparted in the form of positional weight matrices, profiles or HMMs. More ambitious projects investigate whole sets of differentially regulated genes for the presence of a shared set of transcription factor-binding sites. The following describes a selection of tools that seem particularly useful to us. Whenever possible, reference is made to tools that do not require local installation, but rather are accessible via the web. Still, the list cannot be exhaustive, and many resources were left out due to space restraints.

In order to retrieve promoter sequences of genes, both the SOURCE server [35,117] and the TRASER tool [118] were found to be helpful. The first dynamically aggregates information about human, mouse or rat genes from a number of databases and collates the information, including EntrezGene (formerly LocusLink) ID, into easily navigable GeneReports. The latter allows the retrieval of genomic sequence based on an EntrezGene ID. To find known transcriptional start sites of a gene of interest, the UCSC or ENSEMBL genome browsers are perhaps the most easily accessible and comprehensive resource nowadays, whereas for a long time, the Eukaryotic Promoter Database [36] was considered the sole definite resource.

With the exception of TESS [37], which can recognise transcription factor-binding sites also by pattern matching, the algorithms discussed in the following use matrix-based models, generalised profiles or HMMs to predict binding sites. Much of the pioneering work in the prediction of transcription factor-binding sites has come from the groups of Stormo, Schneider, Bucher and Werner [26,38-41]. Algorithms such as MatInspector [21] and Match [109] use binding sites defined in TRANSFAC to predict sites in query sequences based on PWMs. Another approach, Delila-Genome, uses an information-theoretic approach and PWMs to allow the analysis of whole genomes for binding sites, and has been recently used in an analysis of PXR/RXR binding sites [42,119]. NHR-scan is an algorithm that was published recently [43,120] as a tool to find binding sites for nuclear hormone receptors. It models the whole two-part binding site of nuclear receptors as an HMM. In contrast, NUBIScan models each of the response element's half-sites separately [20].

Many of the algorithms mentioned above have been developed for quite some time, and have proved their usefulness in a number of projects. However, a problem often encountered with search tools that solely rely on matching matrix or HMM models to a query sequence is a low ability to discriminate between true matches that exhibit activity in *in vitro* tests and random occurrences of sequences resembling a binding site. This is understandable, as these algorithms need to detect short degenerate sequences in long genomic sequences.

Therefore, recent developments have focused on improving the signal to noise ratio in the detection of transcription factor-binding sites. This can be achieved mainly through two approaches: either the binding site to be detected is complemented with additional information, so that the

model becomes more specific, or the sequence to be analysed is filtered before the actual search, so that only stretches of sequence likely to actually contain a binding site are analysed. Evidently, these two approaches for improvement can be used in conjunction, as well.

3.4 Improving the quality of predictions

3.4.1 Making the model more complex

Physiological and pharmacological signals to up- or down-regulate a gene are often mediated by several transcription factors acting in concert. This is reflected by the fact that in many cases, binding sites for these transcription factors are found combined in so-called *cis*-regulatory modules (CRMs), regulatory stretches of DNA ~ 300 – 500 bp long. This allowed for the development of a number of algorithms that search for such CRMs rather than for individual transcription factor-binding sites.

CRM-based algorithms can be broadly divided into two classes: unsupervised algorithms, such as EMCModule [44], ModuleSearcher [45,121] or CREME [46,122], discover CRMs in input sequences, starting from a database of putative transcription factor-binding sites; other algorithms, such as FastM [47], ModelInspector [48], Ahab [49,123], Stubb [50], MScan [51,124] or ClusterBuster [52,125] are scoring putative CRMs based on user-defined CRM models.

Podvinec *et al.* have developed an algorithm called NUBIScan, which also exploits the complexity of a composite model to improve the quality of predictions [20]. This algorithm was designed to detect binding sites of nuclear receptors, such as CAR or PXR, and uses the fact that these receptors bind as heterodimers to two binding sites that are within a few bases of each other to improve the quality of predictions. This algorithm has been used successfully for the detection of functional nuclear receptor-binding sites in sequences as large as 30 kb [53-57].

3.4.2 Filtering the query sequence

Another helpful approach to increase the specificity of results from a binding site search is to narrow down the amount of sequence being analysed. If there is a way to identify parts of the sequence more likely to contain regulatory elements, subsequent searches can concentrate on those parts. This way, false positives are suppressed by leaving out large parts of the sequence; the risk of incurring false negative results is increased, however.

A first selection is often done by focusing on a particular part of a gene's sequence. Most of the time, transcription factor-binding sites are close to the transcriptional start site of the gene, but notably, binding sites are also commonly within introns of the gene they regulate, as well as in the 3' flanking region of genes. Scaffolding/matrix attachment regions (S/MAR) within the chromosome serve as regulatory isolators and provide a natural boundary of the sequence to be analysed [58,59]. Although there has been work on cataloguing and detecting these sites [60,109], their detection remains nontrivial,

and may not be worth the effort, especially as S/MAR sites are usually spaced far apart. In a first approach, analysis is often simply focused on a 5'-extended region of the gene under study, such as the 10 kb leading up to the transcriptional start site, possibly including the first exon. This choice often works astonishingly well. Whereas such a large query sequence will lead to the detection of spurious sites, clusters of sites can, for instance, be tested for function with relative ease *in vitro* [53]. It has to be kept in mind, however, that transcription factors can bind further away from the gene.

Another way of filtering the query sequence is to remove repetitive elements that occur all over the genome. Filtering must be done judiciously, however, as it does not take much to remove a binding site along with the repetitive element near it.

Phylogenetic footprinting has been one of the biggest advances in this field in recent years. This technique is based on the observation that selective pressure acts on different parts of the genome with different rates. The hypothesis underlying the technique is that functional elements within the genome are on average subject to higher selective pressure than nonfunctional parts of the genome, due to the fact that mutations within a functional element are more likely to be detrimental to the organism and, therefore, are selected against more strongly. This should allow the identification of functional elements by performing cross-species comparisons of genome sequences. As a spin-off of genome sequencing programs, novel pairwise alignment algorithms for long sequences sharing only short regions of high similarity have been developed, such as Glass [61], AVID [62,122] or LAGAN [63]. If per cent pairwise identity is calculated in an alignment of the same gene from different organisms, using tools such as PipMaker [64,126], exons as an example of functional elements are typically easily discernible in a plot of this value along the sequence, as the pairwise identity is much higher within coding sequence compared with introns.

In the last years, phylogenetic footprinting has become the method of choice for improving predictions of transcription factor binding. From a sequence of many thousands of bases, phylogenetic footprinting on average retains only 20% as short subsequences of 150 – 200 bp length [65], which subsequently can be analysed for binding sites. This reduction of the sequence to be searched increases the signal to noise ratio, as a lot of 'uninteresting' sequence is prevented from interfering. However, there is evidence that regulatory sites are not necessarily conserved even among closely related species [66-68]. By using phylogenetic footprinting as a filtering step, one may thus risk discarding valid predictions.

A number of algorithms are available on the web, such as rVISTA [69,122], ConSite [70,127], Toucan Workbench [71,121] or oPOSSUM [72], which allow for phylogenetic footprinting of a sequence, followed by analysis for binding sites using positional weight matrices. Most of these algorithms add another step of cross-species validation by default: not only do they constrain the search to conserved regions, but finally only report transcription factor-binding sites whose predictions

overlap in both species. A more visual approach to browsing a gene of choice and examine evolutionary conserved regions is offered by the ECR Browser [73,122]. Additional phylogenetic footprinting resources are available through the Phylofoot portal site [128].

4. Conclusion

The ultimate algorithm that predicts the induction potential of a compound based on its three-dimensional structure and knowledge of the cognate regulatory pathways and genes likely to be induced or repressed remains beyond reach for the foreseeable future. Still, research is steadily progressing, both in the study of receptor–ligand interactions by means of molecular mechanics, docking or quantitative structure–activity relationship methods, and in the field of correctly predicting the effect of transcription factors on cellular regulatory networks.

This review focuses solely on the latter question: which genes will be activated by a particular transcription factor? Naive methods to discover conserved sequences among coregulated genes are valuable in cases in which little is known about likely mediators. Despite high noise levels, recent algorithms can extract regulatory pathways from such experiments, for instance in [74,75].

In cases where a transcription factor set of 'usual suspects' is known, approaches using PWMs or HMMs are well established and the logical choice. The amount of predictions returned by these methods can be overwhelming [76]. Recent advances have, therefore, focused on better postprocessing of the results, be it by identifying and visualising clusters of transcription factors [77] or by filtering for conserved sites, using phylogenetic footprinting to reduce the amount of false positives. This technique has become useful for the study of regulation in disease-relevant organisms only recently, with the increasing availability of metazoan genome sequences.

Two points, however, are crucial in a phylogenetic footprinting study. First, the species to be compared need to be well chosen. Footprinting with two species that are too closely related will not be efficient, as too much non-functional DNA is shared between them, as is the case for mouse and rat. On the other hand, comparison between species too far apart, such as human and the pufferfish *Fugu rubripes* will often not reveal any conserved noncoding regions of interest. Second, one needs to be aware that phylogenetic footprinting selects for regulatory elements common between species. This is desired and necessary in order to reduce the false positive rate, and estimates suggest that 70% of all transcription factor-binding sites are contained within conserved regions [65,66]. This fact may, however, give rise to missed functional regulatory sites whenever regulation is species specific. One example is the upregulation of murine CYP7A transcription in response to oxysterols, mediated by the LXR. In man, no such regulation is seen, as the regulatory element is absent in CYP7A1, the human orthologous gene.

It cannot be denied that familiarising oneself with bioinformatics methods currently available for the analysis of response elements resembles an uphill march at good times, and orienteering in the fog at bad times. Still, time invested in understanding the methodology and particular strengths and weaknesses of the different approaches is time well spent, as knowledge thus gained will make up time by allowing the combination of fast *in silico* prediction with a targeted *in vitro* follow-up. It is the authors' hope that the present text, in particular its references to available algorithms, will provide a road map to assist in research and keep orienteering to a minimum.

5. Expert opinion

The discovery of response elements with the sequence of target genes is an important step in the analysis of the exact mechanisms of that gene's regulation. The definition of a short subsequence (< 300 bp) mediating regulation enables the pursuit of numerous experimental approaches that allow the further characterisation of this response, such as mutagenesis experiments, DNase I footprinting, DNA-affinity purification, EMSA or transactivation assays. These standard experiments allow the thorough characterisation of a binding site of interest, in terms of what nucleotides are needed for interaction, which proteins it can interact with, and whether protein binding patterns change dependent on treatment condition. All of the above methods, however, require a sufficiently short response element sequence.

The experimental definition of short response elements is both work intensive and time consuming. Classical approaches, such as DNase I hypersensitivity assays or dissection of a large regulatory region into smaller subfragments responsive in reporter gene assays, are tedious. Here, any computational approach offering a possible shortcut is welcome. Although it is unrealistic to postulate that *in silico* prediction of response elements can replace experimental evidence at the current state of knowledge about DNA-protein interactions, computational analysis can lead to the selection of a manageable number of likely sites of regulation that can then be tested *in vitro* rapidly and efficiently. If the predictions prove true, a significant amount of time and work has been saved, and if no activity is found to be mediated by the predicted sites, little time was lost. Therefore, any approach that has a reasonable chance of predicting functional sites even with a fairly low success ratio is valuable, as it leads to easily and quickly verifiable predictions.

It is impossible to provide an all-encompassing protocol for the analysis of binding sites. We have found it useful to start the analysis with a set of genes, even a small set, that are coexpressed. Next, we find it important to narrow down the search space: try to define candidate receptors and transcription factors based on known regulatory pathways, and try to narrow down the location of likely response elements to conserved elements, doing phylogenetic footprinting. A typical starting point for such studies would be pairwise human-mouse comparisons, with the benefit that assays to test response elements from both species are already well established in the laboratory. When searching for transcription factor-binding sites, also keep in mind that binding sites often cluster into modules, which makes searches more specific.

In the typical use cases, such as the search for transcription factor-binding sites driving the expression of a particular gene of interest, or the search for transcription factor-binding sites responsible for the enhanced expression of a set of coexpressed genes, a shortlist of predicted transcription factor-binding sites will be the successful outcome of computational approaches, once the search space has been sufficiently narrowed down. At this point, predictions need to be verified experimentally. We regularly study such predictions by amplifying the area surrounding the binding site by polymerase chain reaction, and inserting the resulting sequence into a reporter vector. Up to a few dozen predicted sites can be feasibly tested for functionality in this way using reporter gene assays or nuclear receptor transactivation assays. Sites showing response can then further be characterised by site-directed mutagenesis or electrophoretic mobility shift assays. All of these experimental techniques, however, have a notable shortcoming, as they test the function of response elements within the context of nonchromatinised DNA. Further techniques, such as *in vitro* footprinting or chromatin immunoprecipitation, can be subsequently used to elucidate the activity of elements in the context of chromatin.

To conclude, we believe that bioinformatics and wet-lab science can only unleash their true potential when they work together. On one hand, all bioinformatics methods rely in a more or less direct way on knowledge derived from *in vitro* or *in vivo* experiments, and on the other hand, predictions obtained with algorithms and models need to be verified experimentally in the laboratory. There must be constant feedback between the two disciplines, and a sensible fusion of both methodologies can achieve greater leaps than any one by itself.

Bibliography

Papers of special note have been highlighted as of interest (•) to readers.

1. DICKINS M: Induction of cytochromes P450. *Curr. Top. Med. Chem.* (2004) **4**(16):1745-1766.
- **Drug induction seen from the medicinal chemistry perspective.**
2. KLAASSEN CD, SLITT AL: Regulation of hepatic transporters by xenobiotic receptors. *Curr. Drug Metab.* (2005) **6**(4):309-328.
- **Drug induction seen from the transporter perspective.**
3. HANDSCHIN C, MEYER UA: Induction of drug metabolism: the role of nuclear receptors. *Pharmacol. Rev.* (2003) **55**(4):649-673.
- **Drug induction seen from the nuclear receptor perspective.**
4. BIRNEY E, ANDREWS D, BEVAN P *et al.*: Ensembl 2004. *Nucleic Acids Res.* (2004) **32**:D468-D470.
- **User-friendly access to a wealth of genome information.**
5. KENT WJ, SUGNET CW, FUREY TS *et al.*: The human genome browser at UCSC. *Genome Res.* (2002) **12**(6):996-1006.
6. ODOM DT, ZIZLSPERGER N, GORDON DB *et al.*: Control of pancreas and liver gene expression by HNF transcription factors. *Science* (2004) **303**(5662):1378-1381.
7. HOOD L, HEATH JR, PHELPS ME, LIN B: Systems biology and new technologies enable predictive and preventative medicine. *Science* (2004) **306**(5696):640-643.
8. REYES H, REISZ-PORSZASZ S, HANKINSON O: Identification of the Ah receptor nuclear translocator protein (Arnt) as a component of the DNA binding form of the Ah receptor. *Science* (1992) **256**(5060):1193-1195.
9. GLASS CK: Some new twists in the regulation of gene expression by thyroid hormone and retinoic acid receptors. *J. Endocrinol.* (1996) **150**(3):349-357.
10. TIRONA RG, KIM RB: Nuclear receptors and drug disposition gene regulation. *J. Pharm. Sci.* (2005) **94**(6):1169-1186.
11. PASCUSI JM, GERBAL-CHALOIN S, DROCOURT L *et al.*: Cross-talk between xenobiotic detoxication and other signalling pathways: clinical and toxicological consequences. *Xenobiotica* (2004) **34**(7):633-664.
12. HANDSCHIN C, PODVINEC M, AMHERD R *et al.*: Cholesterol and bile acids regulate xenosensor signaling in drug-mediated induction of cytochromes P450. *J. Biol. Chem.* (2002) **277**(33):29561-29567.
13. WRAY GA, HAHN MW, ABOUHEIF E *et al.*: The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* (2003) **20**(9):1377-1419.
14. RODRIGUEZ-TRELLES F, TARRIO R, AYALA FJ: Evolution of *cis*-regulatory regions versus codifying regions. *Int. J. Dev. Biol.* (2003) **47**(7-8):665-673.
15. YUH CH, BOLOURI H, DAVIDSON EH: Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* (1998) **279**(5358):1896-1902.
16. JUGE-AUBRY C, PERNIN A, FAVEZ T *et al.*: DNA binding properties of peroxisome proliferator-activated receptor subtypes on various natural peroxisome proliferator response elements. Importance of the 5'-flanking region. *J. Biol. Chem.* (1997) **272**(40):25252-25259.
17. ROULET E, BUSSO S, CAMARGO AA *et al.*: High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.* (2002) **20**(8):831-835.
18. LAVORGNA G, BONCINELLI E, WAGNER A, WERNER T: Detection of potential target genes *in silico*? *Trends Genet.* (1998) **14**(9):375-376.
19. FRECH K, QUANDT K, WERNER T: Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci.* (1997) **22**(3):103-104.
20. PODVINEC M, KAUFMANN MR, HANDSCHIN C, MEYER UA: NUBIScan, an *in silico* approach for prediction of nuclear receptor response elements. *Mol. Endocrinol.* (2002) **16**(6):1269-1279.
21. QUANDT K, FRECH K, KARAS H, WINGENDER E, WERNER T: MatFind and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* (1995) **23**(23):4878-4884.
22. HERTZ GZ, STORMO GD: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* (1999) **15**(7-8):563-577.
23. SHANNON CE: A mathematical theory of communication. *Bell System Technical Journal* (1948) **27**:379-423 and 623-656.
24. SCHNEIDER TD, STEPHENS RM: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* (1990) **18**(20):6097-6100.
25. CROOKS GE, HON G, CHANDONIA JM, BRENNER SE: WebLogo: a sequence logo generator. *Genome Res.* (2004) **14**(6):1188-1190.
26. BUCHER P, KARPLUS K, MOERI N, HOFMANN K: A flexible motif search technique based on generalized profiles. *Comput. Chem.* (1996) **20**(1):3-23.
27. EDDY SR: Profile hidden Markov models. *Bioinformatics* (1998) **14**(9):755-763.
28. WINGENDER E, CHEN X, HEHL R *et al.*: TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* (2000) **28**(1):316-319.
29. SANDELIN A, ALKEMA W, ENGSTROM P, WASSERMAN WW, LENHARD B: JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* (2004) **32**:D91-D94.
- **A well-curated, open database of transcription factor-binding sites.**
30. KOLCHANOV NA, IGNATIEVA EV, ANANKO EA *et al.*: Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res.* (2002) **30**(1):312-317.
31. BAILEY TL, ELKAN C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* (1994) **2**:28-36.
32. THOMPSON W, ROUCHKA EC, LAWRENCE CE: Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.* (2003) **31**(13):3580-3585.
33. LIU XS, BRUTLAG DL, LIU JS: An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* (2002) **20**(8):835-839.
34. RIGOUTSOS I, FLORATOS A: Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* (1998) **14**(1):55-67.
35. DIEHN M, SHERLOCK G, BINKLEY G *et al.*: SOURCE: a unified genomic resource of functional annotations, ontologies, and

- gene expression data. *Nucleic Acids Res.* (2003) **31**(1):219-223.
36. SCHMID CD, PRAZ V, DELORENZI M, PERIER R, BUCHER P: The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res.* (2004) **32**:D82-D85.
37. SCHUG JO, GC: *TESS: Transcription Element Search Software on the WWW*. (Technical Report CBIL-TR-1997-1001-v0.0) Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania, USA (1997).
38. SCHNEIDER TD, STORMO GD, GOLD L, EHRENFEUCHT A: Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* (1986) **188**(3):415-431.
39. STORMO GD: DNA binding sites: representation and discovery. *Bioinformatics* (2000) **16**(1):16-23.
40. SCHNEIDER TD: Consensus sequence *Zen. Appl. Bioinformatics* (2002) **1**(3):111-119.
41. WERNER T: Computer-assisted analysis of transcription control regions. MatInspector and other programs. *Methods Mol. Biol.* (2000) **132**:337-349.
42. VYHLIDAL CA, ROGAN PK, LEEDER JS: Development and refinement of pregnane X receptor (PXR) DNA binding site model using information theory: insights into PXR-mediated gene regulation. *J. Biol. Chem.* (2004) **279**(45):46779-46786.
43. SANDELIN A, WASSERMAN WW: Prediction of nuclear hormone receptor response elements. *Mol. Endocrinol.* (2005) **19**(3):595-606.
44. GUPTA M, LIU JS: *De novo cis*-regulatory module elicitation for eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* (2005) **102**(20):7079-7084.
45. AERTS S, VAN LOO P, THUIS G, MOREAU Y, DE MOOR B: Computational detection of *cis*-regulatory modules. *Bioinformatics* (2003) **19**(Suppl. 2):II5-II14.
46. SHARAN R, BEN-HUR A, LOOTS GG, OVCHARENKO I, CREME: *Cis*-Regulatory Module Explorer for the human genome. *Nucleic Acids Res.* (2004) **32**:W253-W256.
47. KLINGENHOFF A, FRECH K, QUANDT K, WERNER T: Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* (1999) **15**(3):180-186.
48. FRECH K, DANESCU-MAYER J, WERNER T: A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J. Mol. Biol.* (1997) **270**(5):674-687.
49. RAJEWSKY N, VERGASSOLA M, GAUL U, SIGGIA ED: Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* (2002) **3**(1):30.
50. SINHA S, VAN NIMWEGEN E, SIGGIA ED: A probabilistic method to detect regulatory modules. *Bioinformatics* (2003) **19**(Suppl. 1):i292-i301.
51. ALKEMA WB, JOHANSSON O, LAGERGREN J, WASSERMAN WW: MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.* (2004) **32**:W195-W198.
52. FRITH MC, LI MC, WENG Z: Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* (2003) **31**(13):3666-3668.
53. PODVINEC M, HANDSCHIN C, LOOSER R, MEYER UA: Identification of the xenosensors regulating human 5-aminolevulinate synthase. *Proc. Natl. Acad. Sci. USA* (2004) **101**(24):9127-9132.
54. VAISANEN S, DUNLOP TW, SINKKONEN L, FRANK C, CARLBERG C: Spatio-temporal activation of chromatin on the human CYP24 gene promoter in the presence of 1 α ,25-dihydroxyvitamin D3. *J. Mol. Biol.* (2005) **350**(1):65-77.
55. GENOLET R, KERSTEN S, BRAISSANT O *et al.*: Promoter rearrangements cause species-specific hepatic regulation of the glyoxylate reductase/hydroxypyruvate reductase gene by the peroxisome proliferator-activated receptor- α . *J. Biol. Chem.* (2005) **280**(25):24143-24152.
56. GUPTA RK, VATAMANIUK MZ, LEE CS *et al.*: The MODY1 gene HNF-4 α regulates selected genes involved in insulin secretion. *J. Clin. Invest.* (2005) **115**(4):1006-1015.
57. JUNG D, FANTIN AC, SCHEURER U, FRIED M, KULLAK-UBLICK GA: Human ileal bile acid transporter gene ASBT (SLC10A2) is transactivated by the glucocorticoid receptor. *Gut* (2004) **53**(1):78-84.
58. VON STERNBERG R, SHAPIRO JA: How repeated retroelements format genome function. *Cytogenet. Genome Res.* (2005) **110**(1-4):108-116.
59. GOETZE S, BAER A, WINKELMANN S *et al.*: Performance of genomic bordering elements at predefined genomic loci. *Mol. Cell. Biol.* (2005) **25**(6):2260-2272.
60. FRISCH M, FRECH K, KLINGENHOFF A *et al.*: *In silico* prediction of scaffold/matrix attachment regions in large genomic sequences. *Genome Res.* (2002) **12**(2):349-354.
61. BATZOGLOU S, PACTHER L, MESIROV JP, BERGER B, LANDER ES: Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* (2000) **10**(7):950-958.
62. BRAY N, DUBCHAK I, PACTHER L: AVID: A global alignment program. *Genome Res.* (2003) **13**(1):97-102.
63. BRUDNO M, DO CB, COOPER GM *et al.*: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* (2003) **13**(4):721-731.
64. SCHWARTZ S, ZHANG Z, FRAZER KA *et al.*: PipMaker – a web server for aligning two genomic DNA sequences. *Genome Res.* (2000) **10**(4):577-586.
65. LENHARD B, SANDELIN A, MENDOZA L *et al.*: Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* (2003) **2**(2):13.
66. DERMITZAKIS ET, CLARK AG: Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* (2002) **19**(7):1114-1121.
67. COSTAS J, CASARES F, VIEIRA J: Turnover of binding sites for transcription factors involved in early *Drosophila* development. *Gene* (2003) **310**:215-220.
68. EMBERLY E, RAJEWSKY N, SIGGIA ED: Conservation of regulatory elements between two species of *Drosophila*. *BMC Bioinformatics* (2003) **4**:57.
69. LOOTS GG, OVCHARENKO I: rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.* (2004) **32**:W217-W221.
70. SANDELIN A, WASSERMAN WW, LENHARD B: ConSite: web-based

- prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.* (2004) **32**:W249-W252.
71. AERTS S, VAN LOO P, THIJIS G *et al.*: TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.* (2005) **33**:W393-W396.
 72. HO SUI SJ, MORTIMER JR, ARENILLAS DJ *et al.*: oPOSSUM: identification of over-represented transcription factor binding sites in coexpressed genes. *Nucleic Acids Res.* (2005) **33**(10):3154-3164.
 73. OVCHARENKO I, NOBREGA MA, LOOTS GG, STUBBS L: ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res.* (2004) **32**:W280-W286.
 74. MOOTHA VK, HANDSCHIN C, ARLOW D *et al.*: Err- α and Gabpa/b specify PGC-1 α -dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proc. Natl. Acad. Sci. USA* (2004) **101**(17):6570-6575.
 75. LI W, MEYER CA, LIU XS: A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* (2005) **21**(Suppl. 1):i274-i282.
 76. VAVOURI T, ELGAR G: Prediction of cis-regulatory elements using binding site matrices – the successes, the failures and the reasons for both. *Curr. Opin. Genet. Dev.* (2005) **15**(4):395-402.
 77. DI CARA A, SCHMIDT K, HEMMINGS BA, OAKELEY EJ: PromoterPlot: a graphical display of promoter similarities by pattern recognition. *Nucleic Acids Res.* (2005) **33**:W423-W426.
- Websites**
- Websites of special note have been highlighted as of interest (•) to readers.
101. <http://www.genomesonline.org> GOLD Genomes OnLine database (2005).
 102. <http://www.genome.gov/10001356> Remarks made by the President, Prime Minister Tony Blair of England (via satellite), Dr Francis Collins, Director of the National Human Genome Research Institute, and Dr Craig Venter, President and Chief Scientific Officer, Celera Genomics Corporation, on the Completion of the first survey of the entire Human Genome Project (2000).
 103. <http://www.ensembl.org/index.html> Ensembl Genome browser (2005).
 104. <http://genome.ucsc.edu/> Genome browser (2005).
 105. <http://www.expasy.org/tools/scanprosite/scanprosite-doc.html> ScanProsite tool manual (2004).
 106. <http://weblogo.berkeley.edu/> WebLogo (2005).
 107. <http://www.isrec.isb-sib.ch/ftp-server/pftools/> Download of the pftools package (2005).
 108. <http://hmmer.wustl.edu/> HMMER: sequence analysis using profile hidden Markov models (2005).
 109. <http://www.gene-regulation.com/> Gene Regulation portal site (2005).
 - **Partially commercial site, provides access to a number of databases and tools to analyse gene regulation.**
 110. <http://srs.ebi.ac.uk/> SRS Release 7.1.3.1 (2005).
 111. <http://jaspar.cgb.ki.se/> The JASPAR database (2004).
 112. <http://www.mgs.bionet.nsc.ru/mgs/gnw/trrd/> TRRD – Transcription Regulatory Regions Database (2005).
 113. <http://meme.sdsc.edu/> MEME website (2005).
 114. <http://bayesweb.wadsworth.org/gibbs/gibbs.html> The Gibbs Motif Sampler homepage (2005).
 115. <http://seqmotifs.stanford.edu> A suite of web-based programs to search for regulatory motifs in prokaryotes and eukaryotes (2004).
 116. <http://cbcsrv.watson.ibm.com/Tspd.html> TEIRESIAS: Sequence Pattern Discovery (2005).
 117. <http://source.stanford.edu> SOURCE search (2005).
 118. <http://genome-www6.stanford.edu/cgi-bin/Traser/traser> TRASER – Transcript Sequence Retrieval (2005).
 119. <http://r.faculty.umkc.edu/roganp/Information/delgen.html> Delila-Genome webpage (2005).
 120. http://mordor.cgb.ki.se/cgi-bin/NHR-scan/nhr_scan.cgi NHR-Scan (2005).
 121. <http://www.esat.kuleuven.ac.be/~saerts/software/toucan.php> TOUCAN – regulatory sequence analysis (2005).
 122. <http://www.dcode.org/> Comparative Genomics Center at Lawrence Livermore National Laboratory (2005).
 - **Portal to a large number of comparative genomics tools.**
 123. <http://gaspard.bio.nyu.edu/Ahab.html> Ahab (2003).
 124. <http://mscan.cgb.ki.se/cgi-bin/MSCAN>
 125. <http://zlab.bu.edu/cluster-buster/>
 126. <http://pipmaker.bx.psu.edu/cgi-bin/pipmaker> PipMaker and MultiPipMaker (2005).
 127. <http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/> CONSITE (2005).
 128. <http://www.phylofoot.org/> Tools for phylogenetic footprinting (2005).
 - **Portal to a number of phylogenetic footprinting resources.**
 129. <http://perldoc.perl.org/perlretut.html> perlretut – Perl regular expressions tutorial (2000).
- Affiliation**
Michael Podvinec^{†1} PhD & Urs A Meyer² MD
[†]Author for correspondence
¹Swiss Institute of Bioinformatics and Biozentrum, University of Basel, Klingelbergstrasse 50-70, CH-4056 Basel, Switzerland
Tel: +41 61 267 15 83; Fax: +41 61 267 15 84;
E-mail: michael.podvinec@unibas.ch
²Biozentrum, University of Basel, Klingelbergstrasse 50-70, CH-4056 Basel, Switzerland
Tel: +41 61 267 22 20; Fax: +41 61 267 22 08;
E-mail: urs-a.meyer@unibas.ch