BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 5. 509–524
Advance Access publication June 4, 2009
doi:10.1093/bib/bbp025

Computational methods for the detection of *cis*-regulatory modules

Peter Van Loo and Peter Marynen

Submitted: 2nd March 2009; Received (in revised form): 19th April 2009

Abstract

Metazoan transcription regulation occurs through the concerted action of multiple transcription factors that bind co-operatively to *cis*-regulatory modules (CRMs). The annotation of these key regulators of transcription is lagging far behind the annotation of the transcriptome itself. Here, we give an overview of existing computational methods to detect these CRMs in metazoan genomes. We subdivide these methods into three classes: CRM scanners screen sequences for CRMs based on predefined models that often consist of multiple position weight matrices (PWMs). CRM builders construct models of similar CRMs controlling a set of co-regulated or co-expressed genes. CRM genome screeners screen sequences or complete genomes for CRMs as homotypic or heterotypic clusters of binding sites for any combination of transcription factors. We believe that CRM scanners are currently the most advanced methods, although their applicability is limited. Finally, we argue that CRM builders that make use of PWM libraries will benefit greatly from future advances and will prove to be most instrumental for the annotation of regulatory regions in metazoan genomes.

Keywords: *transcription regulation; cis-regulatory modules; genome annotation; regulatory regions; computational CRM detection*

INTRODUCTION

In contrast to the available knowledge about genes, the annotation of gene regulatory regions in metazoan genomes is far from complete. One of the most essential mechanisms used to control gene expression is regulation of transcription initiation. Binding of transcription factors to DNA plays a key role in transcription regulation. The sequences to which these transcription factors bind are very short (about 5–12 base pairs) and show considerable variability. Hence, potential transcription factor binding sites are observed in abundance in large genomic sequences, few of which are really functional. Wasserman and Sandelin termed this the 'futility theorem' [1], stating that the grand majority of predicted transcription factor binding sites is non-functional.

Transcriptional regulatory sequences are often composed of multiple binding sites for multiple transcription factors. Through the concerted binding of a specific combination of transcription factors and co-factors, gene transcription can be tightly controlled [2, 3]. Because of their modular composition and their action in *cis*, these sequences are called *cis*-regulatory modules (CRMs). Although the degeneracy of transcription factor binding sites strongly complicates the reliable detection of these CRMs, it is exactly this binding site clustering that helps to overcome the futility theorem, justifying the development of computational methods to detect CRMs.

A second biologically inspired principle to try to overcome the futility theorem is the use of sequence conservation between related species [4]. These comparative genomics are motivated by the observation that functional sequences accumulate fewer mutations during evolution than non-functional sequences.

Corresponding author. Peter Van Loo, Department of Human Genetics, VIB and University of Leuven, Herestraat 49, Box 602, B-3000 Leuven, Belgium. Tel: +32 16 33 01 44; Fax: +32 16 34 71 66; E-mail: peter.vanloo@med.kuleuven.be

Peter Van Loo, is a Postdoctoral Researcher at the Department of Molecular and Developmental Genetics, VIB and the Department of Human Genetics, University of Leuven, Belgium. His research interests lie in gene regulatory bioinformatics, high-throughput technologies and cancer research.

Peter Marynen, is a Full Professor at the Department of Molecular and Developmental Genetics, VIB and the Department of Human Genetics, University of Leuven, Belgium. His research interests lie in human genomics, mental retardation and cancer.

Genes that are co-regulated or co-expressed in a specific process can be hypothesized to share regulatory signals. Thus, by searching for similar CRMs in co-regulated genes, one might expect a further improvement of the signal-to-noise ratio.

Finally, transcription factor binding sites can be modeled effectively by position weight matrices (PWMs), and databases of PWMs have been constructed [5, 6]. Although at the moment, these PWM libraries are largely incomplete and often noisy, the search for instances of known motifs is a considerably less complex problem than the search for *de novo* motifs [7].

Despite the use of these biologically inspired principles, the computational detection of functional regulatory sequences in metazoan genomes remains a formidable challenge. In this review, we focus on computational methods to detect *cis*-regulatory modules, and on how the above biological inspired principles are utilized by these *in silico* methods.

The existing CRM detection approaches can be classified in three conceptually different classes, based on the specific aims of the methods (Figure 1):

- Methods that scan sequences or complete genomes for CRMs based on a predefined model. These approaches aim to identify CRMs that contain binding sites for a specific combination of PWMs. Hence, they make use of both libraries of known motifs and of clustering of binding sites (for a specific and focused combination of transcription factors). We call these methods CRM scanners.
- Methods that look for similar CRMs in a set of co-regulated or co-expressed genes. These approaches construct or select a combination of PWMs for which binding sites can be found in the putative regulatory regions of some or all of the given co-regulated genes. These methods combine binding site clustering with the assumption that similar expression patterns are controlled by similar regulatory elements. We call these CRM builders. We subdivide this class into methods that construct their own motifs and methods that make use of libraries of PWMs.
- Methods that screen sequences or complete genomes for CRMs as clusters of binding sites for any combination of transcription factors. These methods do not require a predefined model or a predefined set of PWMs, but instead they look for clusters of binding sites for any combination of PWMs. We call these CRM genome screeners. These CRM genome screeners make only few assumptions regarding the CRMs they aim to detect. They are the most generally applicable methods: they use PWM libraries to find homotypic or heterotypic transcription factor binding site clusters.

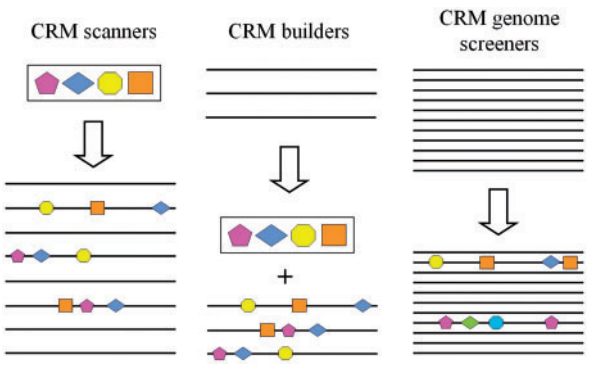

Figure 1: Three different types of *cis*-regulatory module detection algorithms. CRM scanners scan the genome for a user-defined combination of PWMs. CRM builders use a set of co-regulated genes and construct a model of similar CRMs (mostly a combination of PWMs) controlling these genes. CRMs on the given co-regulated genes are predicted as well. CRM genome screeners detect CRMs as homotypic and/or heterotypic clusters, making no assumptions on which transcription factors are involved.

All classes contain early approaches that do not use comparative genomics, and late approaches that have incorporated some measure of evolutionary sequence conservation.

CRM SCANNERS

CRM scanners scan for sequences that satisfy a strictly defined CRM model, often a combination of PWMs (Figure 1). As such, they aim to detect CRMs involved in specific and well-studied processes: the user is asked to supply a rigorous definition of what he is searching for, in the form of PWMs for a set of transcription factors working co-operatively in the process. Because of this strict process definition, the user inherently defines (some aspects of) the expression pattern that the detected CRMs should drive, and this prediction is often used in the validation of these tools.

The properties of the different CRM scanners are outlined in Table 1. These methods commonly

Table 1: Different CRM scanners

Algorithm	Input	Parameters	Principle	Validation	Comments	Availability
LRA [31, 32]	(i) Training set of known CRMs (with identical length); (ii) negative training set; (iii) set of PWMs; (iv) genomic sequence	(i) Score threshold	Logistic regression analysis: model the (probability of) occurrence of CRMs as a function of the transcription factor binding site scores using multivariate logistic regression	(i) Skeletal muscle: 66% sensitivity on training set, 60% sensitivity on test set, one prediction every 32 kb; (ii) liver: 62% sensitivity on training set, 50% in complete jackknife analysis, one prediction per 35 kb	(i) First to show the principle; (ii) no direct comparative genomics, but they do use it as a second step screening strategy	Available in any statistical package
Cister [33]	(i) Set of PWMs; (ii) a sequence	(i) Transcription factor binding site (TFBS) detection threshold; (ii) average distance between TFBSs; (iii) average number of TFBSs; (iv) average distance between CRMs; (v) window size for local background model	Hidden Markov model (HMM)	(i) Regulatory targets of LSF (human): sensitivity: 67%, one prediction every 33 kb; (ii) skeletal muscle: performance comparable to LRA	Output is difficult to interpret	Available as an online tool
Ahab [26]	(i) Putative regulatory sequences of the whole genome; (ii) set of PWMs	(i) Window size; (ii) window step size	Computes via maximum likelihood the probability that the window sequence is made up by sampling from the known PWMs or background (for each window); overlap is allowed and multiple weak instances are taken into account (since all possible segmentations in binding sites are considered)	(i) Body patterning of the *Drosophila* embryo (8 PWMs): 146 CRMs are found in the genome, including 17 of 27 known CRMs, estimated false positive rate is 50%; (ii) [34]: Ahab predictions on the *Drosophila* segmentation network were experimentally validated by reporter constructs: 13 of 16 novel predictions drove expression in a correct pattern	(i) More or less the algorithm of choice (if comparative genomics not available); (ii) claimed to work also when PWMs are defined by Gibbs sampling; (iii) no comparative genomics	Code available upon request
(e)CIS-ANALYST [35, 36]	(i) DNA database; (ii) set of binding sites of cooperatively working transcription factors	(i) Window size; (ii) cut-off score per site (*P*-value); (iii) minimum number of sites	Counts number of sites scoring above threshold; if this number is higher than the minimum number of sites asked, a CRM prediction is returned	CIS-ANALYST: Bcd, Cad, Hb, Kr and Kni in *Drosophila* finds 9 known CRMs and 22 novel predictions (augmented to 28 by also looking for Bcd, Hb, Kr and Kni), 6 of those were positive; eCIS-ANALYST was constructed based on the results	(i) Very simple, but very good performance (on the *Drosophila* segmentation network, although the choice of parameters was dictated by sensitivity/specificity for finding known CRMs)	Available as an online tool

(continued)

Table I: Continued

Algorithm	Input	Parameters	Principle	Validation	Comments	Availability
[37]	–	–	Looks for combinations of two Mad, two Tin, two Twi, two Pnt and one dTCF binding site, derived from *Drosophila* eve dorsal mesodermal enhancer, within a 500 base pair window	1 of 33 predicted enhancers, one was experimentally validated (and fully characterized) and shown to function in a similar way as the eve enhancer; some others were tested by reporter assays but shown to be non-functional	Not a general approach, but applied to a specific example	A perl script is available online
COMET [38]	(i) Set of sequences to search; (ii) set of PWMs	(i) Gap penalty (expected average distance between motifs); (ii) window width for local nucleotide frequency background model	Add PWM scores, use gap penalty for spacer sequences (in fact: HMM); statistics: log likelihood ratio of observing the data given a model of *cis*-element clusters versus a model of background DNA	(i) Promoters regulated by LSF (in combination with SpI, Ets-1 and the TATA box), muscle (Mef2, Myf, SRF, Tef, SpI) (human); (ii) comparison with Cister and LRA: performance is comparable	(i) E-value per CRM (and first to do this); (ii) construction of a model of background DNA is not straightforward; (iii) no comparative genomics	(i) Downloadable executable; (ii) online tool
SCORE [39]	(i) A (whole-genome) sequence to scan; (ii) a consensus sequence	None	Detect over-representation of binding sites of one particular transcription factor in differently sized windows	Applied to Su(H) sites in *Drosophila*: one prediction was successfully validated in the lab	Only homotypic clusters	None stated
Cluster-Buster [40]	(i) Set of sequences to search; (ii) set of PWMs	(i) Gap penalty (in fact: expected average distance between motifs); (ii) window width for (local nucleotide frequency) background model	Similar to COMET	Validated using Gene Ontology term enrichment in (i) muscle and (ii) LPS stimulation [41]	–	(i) Downloadable executable and (ii) online tool
MCAST [42]	(i) DNA database; (ii) set of PWMs	(i) *P*-value cutoff (for a single transcription factor binding site); (ii) maximum gap length; (iii) gap penalty	HMM	(i) Simulated data; (ii) real data in human and *Drosophila*; compared with COMET: similar but slightly better performance	(i) *E*-value per module; (ii) sensitive to the setting of its parameters; (iii) no comparative genomics	Web site is given, but no longer online

ModuleScanner [12]	(i) A set of genomic sequences or conserved non-coding sequences; (ii) a set of PWMs	(i) Max CRM size; (ii) overlap; (iii) penalization	Looks for combination of PWMs gives the highest score (sum of binding energies)	(i) *In silico*: human cell-cycle PWM set predicted by ModuleSearcher validated by Gene Ontology; (ii) *in vitro* (human; [19]): CRMs in upregulated HL-60 cells	–	(i) Available on request; (ii) integrated in Toucan [43]
MSCAN [44]	(i) A set of transcription factor binding profiles (PWMs) and (ii) a sequence	(i) Significance threshold (for single PWM hits); (ii) window size; (iii) maximum number of motifs in a CRM	Looks for significant PWM hit combinations; a *P*-value is assigned to each binding site, and these are later combined to a CRM score (two options: minimum *P*-value or product of *P*-values); this CRM score is then fitted to a statistical distribution to derive a *P*-value	(i) Liver (66% sensitivity, 1 putative CRM detected every 23 kb); (ii) skeletal muscle (66% sensitivity; one putative CRM detected every 15 kb); comparison to LRA, Cister and COMET (slightly better performance)	–	Available as an online tool
Stubb [45,46]	(i) Set of sequences (or a full genome) of one or more species; (ii) set of PWMs	(i) Window length; (ii) window step size; (iii) background model	Based on Ahab, with two modifications: (i) correlation between factors is modeled (e.g. factor A preferentially follows factor B) and (ii) comparative genomics: sequence conservation in multiple species is incorporated (by counting scores in aligned blocks in both species)	(i) Synthetic sequences and yeast toy example; (ii) gap gene system of *Drosophila*: all 16 known CRMs are recovered, together with only two novel predictions; (iii) the *Drosophila melanogaster* segmentation network, including *Drosophila pseudoobscura* sequences (*in silico*, using annotated anterior/posterior-segmentation genes)	(i) The multi-species version significantly outperformed the single-species version; (ii) more or less the standard algorithm when an extra species is available	You can request a copy online
PFR-Searcher [10]	(i) Set of similar PFRs (output of PFR-Sampler, Table 2); (ii) a (usually full genome) PFR database	None	(i) A set of PFRs (phylogenetically footprinted non-coding regions) is collected by aligning two (*Drosophila*) genomes; (ii) selection of conserved regions; (iii) Markov chain discrimination (log-likelihood of PFR generated by a "CRM" HMM compared to a "background" HMM)	Set of co-regulated genes centered around 10 *Drosophila* blastoderm genes that are known to share transcription factor binding sites, leave-one-out cross-validation	–	C-code available for download (after license agreement)

(continued)

Table I: Continued

Algorithm	Input	Parameters	Principle	Validation	Comments	Availability
ModuleFinder [47]	(i) Set of transcription factor binding profiles (PWMs); (ii) sequence	(i) Window width; (ii) window step size; (iii) threshold score	For each window, the number of transcription factor binding sites is considered (along with their evolutionary conservation), and the likelihood of observing this is calculated	(i) Skeletal muscle: sensitivity: 96.3%, specificity: 94.4% (although the threshold score is chosen to maximize these values); (ii) compared to LRA, Cister, Comet and MSCAN: performance is better, but none of the other algorithms use comparative genomics; (iii) in [48]: *Drosophila* muscle founder cells	–	Stated to be available for download, but website does not exist
EEL [8]	(i) Two homologous DNA sequences; (ii) set of PWMs	Four parameters that weigh different aspects of the alignment score (can be calculated based on the full genome)	Aligns sequences in the transcription factor binding site domain	*In silico* and *in vivo*: (i) detects all known *Drosophila* eve enhancers; (ii) expression in transgenic mice embryo's (success rate: $\geq$30%)	Novel idea and high performance	Tool available for download
PhylCRM [25]	(i) Orthologous DNA sequences of multiple species; (ii) phylogenetic tree; (iii) set of PWMs	(i) Maximum window size; (ii) different options to take into account evolutionary conservation	Searches for sequence windows that are enriched for evolutionarily conserved clusters of the input motifs; uses the MONKEY scoring scheme [49]	In combination with the Lever framework: human myogenic differentiation: (i) ROC analysis; (ii) *in vitro* validation (luciferase experiments) of a small set of predictions	–	Tool available for download

require a combination of PWMs and a genomic sequence as input, as well as a variable number of parameters. A number of different principles are used to incorporate homotypic and heterotypic clustering of transcription factor binding sites (PWM hits): e.g. counting of occurrences in a specific window, logistic regression analysis to predict the probability of a CRM hit and hidden Markov models (Table 1).

Validation procedures used range from purely *in silico* to extensive *in vivo* studies, although the latter have been performed mostly in *Drosophila* (Table 1). These validations indicate that the methods are useful in practice to detect CRMs in the complete *Drosophila* genome, although it should be kept in mind that only for a very limited number of processes sufficient data is available to construct a combination of PWMs. Although considerable progress has been made, most notably by the incorporation of comparative genomics in multiple methods, detecting CRMs by a genomewide scan in the considerably larger human genome remains a challenge.

We would like to highlight one recent novel approach ''Enhancer Element Locator (EEL)'' [8], which uses alignments of predicted transcription factor binding sites in two species to make CRM predictions. In this method, first the sequences of both species are used to predict binding sites using the given PWMs. In the second step, the sequences themselves are not used anymore, and a Smith–Waterman alignment [9] of the predicted binding sites is performed. Hence, this method uses comparative genomics in a novel way: EEL does not assume that each transcription factor binding site is conserved, but instead makes the assumption that the order of functional binding sites is conserved. In addition, the alignment score function imposes a conservation of the distances between transcription factor binding sites in a CRM. The validation of this method's predictions in the human-mouse system by expression constructs in transgenic mice embryo's showed a success-rate of over 30%, indicating that this method may achieve sufficient sensitivity and specificity levels to annotate CRMs in the human genome.

CRM BUILDERS

CRM builders use a set of co-regulated or co-expressed genes (or their putative regulatory sequences), and aim to detect similar CRMs, assuming that similar expression patterns are driven by similar regulatory elements. Hence, these methods look for recurring signals in the putative regulatory regions of a set of genes with a similar expression pattern or a similar function. They predict the combination of transcription factors (or PWMs) working co-operatively, as well as the CRMs themselves, as combinations of binding sites for these PWMs (Figure 1). The use of co-regulated or co-expressed genes as input implies that these CRM builders look for CRMs with a user-specified function or expression pattern. Hence, for the detected CRMs, a probable function can be predicted.

We sub-classify these CRM builders into two parts: methods that select PWMs from a PWM library and methods that construct their own PWMs. This distinction is apparent in both approach and performance: the construction of PWMs from putative regulatory sequences is a largely unsolved problem, resulting in more complexity and lower performance for methods that construct their own PWMs. Methods that make use of PWM libraries on the other hand are highly sensitive to the quality and completeness of these PWM libraries. Although currently a limiting factor, we believe the rapid development and deployment of new experimental technologies will dissolve this issue in a relatively short term.

CRM builders that build their own PWMs

These methods can be viewed as extensions of approaches detecting single motifs (see [7] and references therein for an overview of motif detection methods), aiming to overcome the limitations of motif detection by incorporating co-operativity. Often these approaches are based on multiple component models, where the singular motifs and their combination are optimized simultaneously or iteratively (Table 2). Two of these methods incorporate comparative genomics: PRF-sampler [10] (*Drosophila melanogaster* and *Drosophila pseudoobscura*) and the Gibbs Module Sampler [11]. This last algorithm extends the Gibbs sampling approach for single-motif detection to combinations of motifs and to motifs conserved in two species. The performance of CRM builders that build their own PWMs is relatively limited (Table 2), in part because motif detection is a complex and unresolved problem [7].

Table 2: Different CRM builders that build their own PWMs

Algorithm	Input	Parameters	Principle	Validation	Comments	Availability
CO-Bind [50]	(i) Set of co-regulated genes and their sequences; (ii) set of background sequences	(i) Maximum distance between binding sites; (ii) algorithm parameters: step-size and decay factor	Gibbs sampling strategy and neural network (perceptron) to find PWMs for two sets of similar binding sites close together	(i) Synthetic data and (ii) four yeast sets (promoters selected for sharing known binding sites): in three cases, both patterns could be identified	Only combinations of up to two factors	Downloadable executable
PFR-Sampler [10]	Set of co-regulated genes	(i) Number of initial hits (default equal to the number of co-regulated genes); (ii) algorithm parameters	(i) A set of PFRs (phylogenetically footprinted non-coding regions) is collected by aligning two (*Drosophila*) genomes; (ii) selection of conserved regions; (iii) algorithm: simulated annealing, using sum of PFR-Searcher scores	Set of co-regulated genes centered around 10 *Drosophila* blastoderm genes that are known to share transcription factor binding sites, leave-one-out cross-validation	–	C-code is downloadable (after license agreement)
[51]	(i) Set of co-regulated genes (and their putative regulatory sequences); (ii) putative regulatory sequences of all genes in the genome	(i) Maximum distance between adjacent motif occurrences; (ii) maximum overlap between binding sites; (iii) minimum number of genes with a CRM; (iv) *P*-value cutoff	Exhaustively tries all combinations of up to four PWMs; ensures that the co-occurring motifs are sparsely distributed throughout the genome	(i) Random sets of genes (negative control); (ii) yeast cell cycle; (iii) *Drosophila* pattern development: finds three motifs (overlap with true motifs is not discussed); predicted CRMs (regions) correspond to known CRMS; (iv) human muscle regulatory regions: finds the three correct PWMs; predicted CRMs (regions) correspond quite well with known regulatory regions	–	Source code is available upon request

CisModule [52]	Sequences of a set of co-regulated genes	(i) Number of PWMs and (ii) module length	(Generative) hierarchical mixture model, consisting of two levels: (i) CRM vs. background and (ii) (within CRM) TFBSs vs. background; iterative algorithm consisting of two steps (similar to Gibbs sampling): (i) given CRM and TFBS positions, estimate parameters, and (ii) given parameters, estimate (sample) CRM and transcription factor binding site positions	(i) Artificial sequences, (ii) three cases of homotypic clustering in *Drosophila*: sensitivity (based on transcription factor binding sites discovered) was 56%, but number of sites was not that small; in all three cases, the correct PWM was found; (iii) (human) muscle-specific regulatory regions: four PWMs were correctly identified, sensitivity was 88%; an analysis was added checking sensitivity to added negative sequences: in 29 positive and 40 negative sequences, 54% of detected CRMs were in the positive sequences	–	Available online
Gibbs Module Sampler [11]	Set of orthologous sequences of a set of co-regulated genes	(i) Maximum number of motifs; (ii) maximum distance between motifs (hardcoded to 100 base pairs)	Gibbs motif sampler extended to (i) find conserved motifs (sampling over aligned pairs of sites) and (ii) find CRMs as combinations of motifs (including neighbor interactions)	(i) Skeletal muscle: four of five motifs were correctly identified; 17 of 20 CRMs were correctly located in 3 kb sequences (50% overlap); TFBSs: sensitivity: 69%, false positive rate: about 35%; (ii) smaller liver case-study: only HNF1 was detected; (iii) comparison to COMET: roughly similar performance, showing that the addition of comparative genomics can offset the extra difficulty of modeling the PWMs	–	None stated

(continued)

Table 2: Continued

Algorithm	Input	Parameters	Principle	Validation	Comments	Availability
EMCMODULE [53]	(i) Putative regulatory sequences of a set of co-regulated genes; (ii) starting set of PWMs (optional)	(i) Prior probability distribution of binding site occurrence; (ii) distribution of distance between motif sites (truncated geometric distribution)	Statistical model to describe CRM structure (HMM) and an evolutionary Monte Carlo motif screening strategy (similar to a genetic algorithm)	(i) *Bacillus subtilis*; (ii) *Drosophila* early development: PWMs were correctly identified for four of the five factors; compared to Gibbs Module Sampler and CisModule: recovered none of the known motifs; (iii) human skeletal muscle: recovered three of the five motifs; when using JASPAR [54] as starting motifs: recovered four of the five motifs	Can work with a starting set of PWMs (although these need to be selected carefully)	Available online
[55]	Sequence data of a set of co-regulated genes (and a negative set)	(i) Window size; (ii) window step size	Three-component model: (i) motif model (PWMs), (ii) module model (CRMs as weighed PWM combinations; models the probability that a sequence window contains a CRM given the binding site occurrences of the motifs in it) and (iii) regulation model (probabilities that the given positive and negative genes are regulated by the CRM); all three models are logistic functions, optimized by an expectation-maximization algorithm	(i) Simulated data; (ii) yeast data (ChIP-chip, genes sets selected for sharing binding sites for two factors): in 11 of 25 sets, CRMs were identified; 7 of 11 PWMs were correct; (iii) human: CRM predictions done on all 381 Gene Ontology categories: 83 CRMs were identified in 71 Gene Ontology categories; of 203 motifs, 54 correspond to known motifs	Assigns a weight to PWMs	None stated

CRM builders that make use of PWM libraries

These methods construct models of similar CRMs controlling co-expressed or co-regulated genes, by essentially selecting the optimal set of PWMs from PWM libraries such as Transfac [5] and Jaspar [6]. CRMs in (a subset of) the given co-regulated genes are concomitantly identified as well.

The different CRM builders that make use of PWM libraries are outlined in Table 3. The number of available algorithms is relatively limited: only two early approaches fall strictly into this category: ModuleSearcher [12, 13] and CREME [14, 15], as well as our recent ModuleMiner [16] algorithm. The MARSMOTIF algorithm [17] has a slightly different focus: it models microarray gene expression as a function of motif content (similar to REDUCE [18]).

Except for the *in vitro* validation of ModuleSearcher (in combination with ModuleScanner) [19] and the extensive *in silico* validation of ModuleMiner [16], these methods have only been validated to a limited extent (Table 3).

In [16], we compare the performance of our ModuleMiner algorithm to several other CRM builders, confirming the limited performance of those that do not make use of PWM libraries. In addition, we believe that novel high-throughput technologies for identifying transcription factor binding profiles, such as chromatin immunoprecipitation [20] and particularly protein-binding microarrays [21], will allow the construction of large and high-quality PWM libraries in the very near future. Indeed, recent studies already report on new PWM collections appearing from the protein-binding microarray technology [22, 23]. Therefore, we hypothesize that CRM builders that make use of PWM libraries will prove to be more useful for the annotation of regulatory regions in the human genome, compared to the methods that build their own PWMs.

Finally, the use of PWM libraries does not necessarily limit these methods to highly studied transcription factors, as databases of predicted PWMs (e.g. [24]) could be used as well. However, none of these methods have yet been tested in this setting.

Complementarity between CRM builders and CRM scanners

CRM scanners and CRM builders are highly complementary. The CRM models constructed by CRM builders can readily be used by CRM scanners to find additional similar CRMs, and additional new target genes of the process under study. In validation experiments of both ModuleSearcher [19] and ModuleMiner [16], this complementarity has been extensively used.

In addition, CRM builders could be constructed as wrappers around CRM scanners: multiple CRM models are used as input of a CRM scanner, and those that perform well on a given set of co-regulated genes can be selected. ModuleMiner is in fact such a wrapper around the Module Scanner algorithm. Similar wrapper approaches around more advanced CRM scanners would be an interesting avenue for further research.

In a recent study [25], this approach has been taken one step further: Lever, a framework around PhylCRM was created that aims to link motifs and motif combinations to gene sets. Lever takes as input CRM scores predicted by phylCRM and a set of gene sets, and then produces a 'GM matrix' from which significant gene set-motif combinations can be selected. This approach is more general than a CRM builder: multiple hypotheses are tested in two dimensions, such that both the CRM model and the co-regulated genes can be co-selected.

CRM GENOME SCREENERS

These methods do not make any assumptions regarding a specific set of transcription factors working in concert. Instead, they look for CRMs as homotypic and/or heterotypic clusters of transcription factor binding sites (Figure 1). The properties of the different CRM genome screeners are summarized in Table 4. In general, these methods require a database of PWMs and a genomic sequence as input. We can subdivide these methods into early approaches (Argos [26] and TraFac [27]) that delivered proof-of-principle but are not aimed at detecting CRMs genomewide, and late approaches (PreMod [28] and Enhancer Element Locator [8]) that aim to make genomewide predictions. The Regulatory Potential method [29, 30] does not strictly aim to detect CRMs, but calculates the regulatory potential as a function of genomic position, based on two- or three-way alignments.

These methods are more general than the CRM scanners and CRM builders: the latter both aim to detect CRMs with a specific function, while CRM

Table 3: Different CRM builders that use PWM libraries

Algorithm	Input	Parameters	Principle	Validation	Comments	Availability
ModuleSearcher [12,13]	(i) Database of PWMs; (ii) sequences of co-regulated genes	(i) Maximum CRM length; (ii) maximum number of PWMs; (iii) penalization	Identifies PWM combinations with maximum sum of scores in the given set of genes; comparative genomics: looks in conserved non-coding regions	(i) *In silico*: human cell cycle, validated using Gene Ontology; (ii) *in vitro*: differential regulation in HL-60 differentiation [19]	—	(i) Available on request; (ii) integrated in Toucan [43]
CREME [14,15]	(i) Database of PWMs; (ii) promoter sequences of a (large) set of (loosely) co-regulated human genes (and orthologous sequences in the mouse and rat genome)	(i) Maximum CRM length; (ii) maximum number of PWMs; (iii) threshold for individual motif hits	(i) Select only single motifs that are overrepresented (compared to a background set of sequences); (ii) filter similar PWMs (by overlap in predicted binding sites); (iii) hashing algorithm to go through all combinations of PWMs and calculate their combined significance (compared to the expected frequency based on the occurrences of their component motifs); (iv) filter similar CRMs	(i) Cell-cycle data, validated using correlation in microarray data; (ii) stress response data, validated using Gene Ontology	(i) Only 1.5 kb 5′ of TSS is tested; (ii) starts with about 500 loosely co-regulated genes	Available as an online tool
MARSMOTIF [17]	(i) Microarray expression data; (ii) set of candidate motifs	Number of maximum interactions allowed (corresponds to the size of CRMs)	Model microarray gene expression as a function of motif content (PWM score), including combinations of motifs using multivariate adaptive regression splines	(i) Simulated data; (ii) yeast cell-cycle, compared to REDUCE [18]; (iii) application to tissue-specific expression modeling [56]: for 56 tissues (GNF Symatlas), 500 positive (tissue-specific) and negative genes were selected and MARSMOTIF was used to build a classifier; significant performance for 45 tissues, although errors were still large	(i) Mostly finds binary interactions; (ii) more focused on finding transcription factors working cooperatively	Software available after license agreement
ModuleMiner [16]	(i) Database of PWMs; (ii) set of co-regulated genes	none	Similar to ModuleSearcher, but identifies the CRMs that are the most discriminative for the given set of co-regulated genes, compared to the rest of the genome	*In silico* (ROC curves): (i) smooth muscle genes; (ii) compared to CRÈME, ModuleSearcher, EMCModule and CisModule, showing better performance; (iii) application to microarray clusters (tissue-specific expression) and developmental gene sets	–	(i) Online tool; (ii) stand-alone version available on request

Table 4: Different CRM genome screeners

Algorithm	Input	Parameters	Principle	Validation	Comments	Availability
Argos [26]	(i) Genomic sequence; (ii) database of PWMs	(i) Window size; (ii) window step size	Looks for overrepresented motifs is a sequence, and combines five different non-overlapping motifs into one score	Predictions over the full genome: false negative rate estimated to be 50%; one prediction per 5 kb	First of its kind	None stated
TraFaC [27]	(i) Two orthologous sequences; (ii) PWM library	None	Looks for clusters of conserved binding sites in one sequence	Very specific case-studies	Leaves the detection of the CRMs to the interpretation of the user	Available as an online tool
Regulatory Potential [29, 30]	Human–mouse(–rat) alignments	None, although many hard-coded choices are made	Collapses alignment alphabet to fewer symbols and uses a higher order HMM (trained on positive vs. negative sequences)	[57]: Applied to a class of erythroid specific genes, including β-globin: sensitivity: 60%, specificity: 60%	Plots regulatory potential as a function of sequence position, hence does not really detect CRMs	Available as a UCSC genome browser track
PreMod [28]	(i) Human–mouse–rat whole-genome alignments; (ii) database of PWMs	(i) Maximum length of CRMs; (ii) score/*P*-value thresholds for TFBSs and CRM detection	Search for statistically significant clusters of (phylogenetically conserved) binding sites for 1–5 transcription factors (PWMs); homotypic clustering is used extensively	(i) Overlap with known CRMs; (ii) ChIP-chip validation for ER and E2F4 binding sites: low performance	–	Full-genome predictions are available
EEL [8]	(i) Two homologous DNA sequences; (ii) database of PWMs	Four parameters that weigh different aspects of the alignment score (can be calculated based on the full genome)	Aligns sequences in the transcription factor binding site domain	Validated as a Type I algorithm	Designed as a Type I method, with Type III potential	Tool and predictions available for download

genome screeners detect CRMs as any homotypic and/or heterotypic clusters of binding sites. Hence, these CRM genome screeners require no prior knowledge (except for a library of PWMs). However, as a consequence, the performance is lower and no inference can be made about the function of the predicted CRMs or about the expression pattern that these predicted CRMs drive.

CONCLUSIONS

We provide an overview of computational methods to detect CRMs in metazoan genomes, subdividing these methods into three classes. CRM scanners screen sequences for CRMs based on predefined models (combinations of PWMs). CRM builders construct models of similar CRMs controlling a set of co-regulated genes. CRM genome screeners screen sequences or complete genomes for CRMs using no assumptions on specific PWM combinations. CRM scanners are currently the most advanced methods, although their applicability is limited to well-studied processes. CRM builders that make use of PWM libraries are expected to benefit greatly from novel technologies that construct these libraries, and we believe these methods will prove to be most instrumental for the annotation of regulatory regions in metazoan genomes.

Key Points

- Computational methods to detect *cis*-regulatory modules (CRM) can be subdivided into three classes.
- CRM scanners screen sequences for CRMs based on predefined models.
- CRM builders construct models of similar CRMs controlling a set of co-regulated or co-expressed genes.
- CRM genome screeners screen sequences for CRMs as homotypic or heterotypic clusters of binding sites.
- We argue that CRM builders that make use of PWM libraries will prove to be most instrumental for the annotation of regulatory regions in metazoan genomes.

Acknowledgements

We would like to thank Stein Aerts for valuable discussion.

FUNDING

PVL is a postdoctoral researcher of the Research Foundation - Flanders (FWO).

References

1. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 2004;**5**:276–87.
2. Davidson EH. *Genomic regulatory systems: development and evolution*. San Diego, CA, USA: Academic Press, 2001.
3. Balmer JE, Blomhoff R. Anecdotes, data and regulatory modules. *Biol Lett* 2006;**2**:431–4.
4. Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* 2003;**4**:251–62.
5. Matys V, Fricke E, Geffers R, *et al*. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003;**31**:374–8.
6. Vlieghe D, Sandelin A, De Bleser PJ, *et al*. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* 2006;**34**: D95–D97.
7. Tompa M, Li N, Bailey TL, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005;**23**:137–44.
8. Hallikas O, Palin K, Sinjushina N, *et al*. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* 2006;**124**: 47–59.
9. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
10. Grad YH, Roth FP, Halfon MS, *et al*. Prediction of similarly acting *cis*-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura. *Bioinformatics* 2004;**20**: 2738–50.
11. Thompson W, Palumbo MJ, Wasserman WW, *et al*. Decoding human regulatory circuits. *Genome Res* 2004;**14**: 1967–74.
12. Aerts S, Van Loo P, Thijs G, *et al*. Computational detection of *cis*-regulatory modules. *Bioinformatics* 2003;**19**(Suppl 2): II5–14.
13. Aerts S, Van Loo P, Moreau Y, *et al*. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics* 2004;**20**:1974–6.
14. Sharan R, Ovcharenko I, Ben-Hur A, *et al*. CREME: a framework for identifying *cis*-regulatory modules in human-mouse conserved segments. *Bioinformatics* 2003; **19**(Suppl 1):i283–91.
15. Sharan R, Ben-Hur A, Loots GG, *et al*. CREME: *cis*-regulatory module explorer for the human genome. *Nucleic Acids Res* 2004;**32**:W253–6.
16. Van Loo P, Aerts S, Thienpont B, *et al*. ModuleMiner – improved computational detection of *cis*-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol* 2008;**9**:R66.
17. Das D, Banerjee N, Zhang MQ. Interacting models of cooperative gene regulation. *Proc Natl Acad Sci USA* 2004;**101**:16234–9.
18. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. *Nat Genet* 2001;**27**:167–71.

19. Aerts S, Lambrechts D, Maity S, *et al*. Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006;**24**: 537–44.
20. Bulyk ML. DNA microarray technologies for measuring protein-DNA interactions. *Curr Opin Biotechnol* 2006;**17**: 422–30.
21. Mukherjee S, Berger MF, Jona G, *et al*. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* 2004;**36**:1331–9.
22. Berger MF, Badis G, Gehrke AR, *et al*. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 2008;**133**:1266–76.
23. Zhu C, Byers K, McCord R, *et al*. High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Res* 2009;**19**:556–66.
24. Xie X, Lu J, Kulbokas EJ, *et al*. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 2005;**434**: 338–45.
25. Warner JB, Philippakis AA, Jaeger SA, *et al*. Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat Methods* 2008;**5**:347–53.
26. Rajewsky N, Vergassola M, Gaul U, *et al*. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* 2002;**3**:30.
27. Jegga AG, Sherwood SP, Carman JW, *et al*. Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res* 2002;**12**:1408–17.
28. Blanchette M, Bataille AR, Chen X, *et al*. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* 2006;**16**:656–68.
29. Elnitski L, Hardison RC, Li J, *et al*. Distinguishing regulatory DNA from neutral sites. *Genome Res* 2003;**13**:64–72.
30. Kolbe D, Taylor J, Elnitski L, *et al*. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res* 2004;**14**:700–7.
31. Wasserman WW, Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 1998;**278**:167–81.
32. Krivan W, Wasserman WW. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res* 2001;**11**:1559–66.
33. Frith MC, Hansen U, Weng Z. Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics* 2001;**17**: 878–89.
34. Schroeder MD, Pearce M, Fak J, *et al*. Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* 2004;**2**:E271.
35. Berman BP, Nibu Y, Pfeiffer BD, *et al*. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci USA* 2002;**99**: 757–62.
36. Berman BP, Pfeiffer BD, Laverty TR, *et al*. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* 2004;**5**:R61.
37. Halfon MS, Grad Y, Church GM, *et al*. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res* 2002;**12**:1019–28.
38. Frith MC, Spouge JL, Hansen U, *et al*. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res* 2002;**30**: 3214–24.
39. Rebeiz M, Reeves NL, Posakony JW. SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc Natl Acad Sci USA* 2002;**99**:9888–93.
40. Frith MC, Li MC, Weng Z. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* 2003; **31**:3666–8.
41. Bluthgen N, Kielbasa SM, Herzel H. Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res* 2005;**33**:272–9.
42. Bailey TL, Noble WS. Searching for statistically significant regulatory modules. *Bioinformatics* 2003; **19**(Suppl 2):ii16–25.
43. Aerts S, Van Loo P, Thijs G, *et al*. TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res* 2005;**33**: W393–6.
44. Johansson O, Alkema W, Wasserman WW, *et al*. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics* 2003;**19**(Suppl 1):i169–76.
45. Sinha S, van NE, Siggia ED. A probabilistic method to detect regulatory modules. *Bioinformatics* 2003;**19**(Suppl 1): i292–301.
46. Sinha S, Schroeder MD, Unnerstall U, *et al*. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics* 2004;**5**:129.
47. Philippakis AA, He FS, Bulyk ML. Modulefinder: a tool for computational discovery of cis regulatory modules. *Pac Symp Biocomput* 2005;519–30.
48. Philippakis AA, Busser BW, Gisselbrecht SS, *et al*. Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells. *PLoS Comput Biol* 2006;**2**:e53.
49. Moses AM, Chiang DY, Pollard DA, *et al*. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol* 2004;**5**:R98.
50. GuhaThakurta D, Stormo GD. Identifying target sites for cooperatively binding factors. *Bioinformatics* 2001;**17**: 608–21.
51. Kreiman G. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res* 2004;**32**:2889–900.
52. Zhou Q, Wong WH. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci USA* 2004;**101**:12114–9.
53. Gupta M, Liu JS. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci USA* 2005;**102**: 7079–84.

54. Sandelin A, Alkema W, Engstrom P, *et al*. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;**32**:D91–4.

55. Segal E, Sharan R. A discriminative model for identifying spatial *cis*-regulatory modules. *J Comput Biol* 2005;**12**: 822–34.

56. Smith AD, Sumazin P, Xuan Z, *et al*. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci USA* 2006;**103**: 6275–80.

57. King DC, Taylor J, Elnitski L, *et al*. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* 2005;**15**:1051–60.