Motif Discovery in Heterogeneous Sequence Data

A. Prakash, M. Blanchette, S. Sinha, and M. Tompa

Pacific Symposium on Biocomputing 9:348-359(2004)

MOTIF DISCOVERY IN HETEROGENEOUS SEQUENCE DATA

A. PRAKASH

Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195-2350 U.S.A.

M. BLANCHETTE

School of Computer Science
McGill University
Montreal, Quebec, Canada H3A 2A7

S. SINHA

Center for Studies in Physics and Biology
The Rockefeller University
New York, NY 10021 U.S.A.

M. TOMPA

Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195-2350 U.S.A.

Abstract

This paper introduces the first integrated algorithm designed to discover novel motifs in *heterogeneous sequence data*, which is comprised of coregulated genes from a single genome together with the orthologs of these genes from other genomes. Results are presented for regulons in yeasts, worms, and mammals.

1 Regulatory Elements and Sequence Sources

An important and challenging question facing biologists is to understand the varied and complex mechanisms that regulate gene expression: how, when, in what cells, and at what rate is a given gene turned on and off? This paper focuses on one important aspect of this challenge, the discovery of novel *binding sites* in DNA (also called *regulatory elements*) for the proteins involved in such gene regulation. This is an important first step in determining which proteins regulate the gene and how.

Until the present, nearly all regulatory element discovery algorithms have focused on what will be called *homogeneous* data sources, in which all the sequence data is of the same type (see Section 1.1). This paper introduces the first integrated algorithm designed to exploit the richer potential of *heterogeneous sequence data*, which is comprised of coregulated genes from a single genome together with the orthologs of these genes from other genomes.

1.1 Regulatory Elements from Homogeneous Data

A number of algorithms have been proposed for the discovery of novel regulatory elements in nucleotide sequences. Most of these try to deduce the regulatory elements by considering the regulatory regions of several (putatively) coregulated genes from a single genome. Such algorithms search for overrepresented motifs in this collection of regulatory regions, these motifs being good candidates for regulatory elements. Some examples of this approach include Bailey and Elkan[1], Brāzma *et al.*[2], Buhler and Tompa[3], Hertz and Stormo[4], Hughes *et al.*[5], Lawrence *et al.*[6], Lawrence and Reilly[7], Rigoutsos and Floratos[8], Rocke and Tompa[9], Sinha and Tompa[10], van Helden *et al.*[11], and Workman and Stormo[12].

An orthogonal approach deduces regulatory elements by considering orthologous regulatory regions of a *single* gene from *multiple* species. This approach has been used in *phylogenetic footprinting* (Tagle *et al.*[13], Loots *et al.*[14]) and *phylogenetic shadowing* (Boffelli *et al.*[15]). The simple premise underlying these comparative approaches is that selective pressure causes functional elements to evolve at a slower rate than nonfunctional sequences. This means that unusually well conserved sites among a set of orthologous regulatory regions are good candidates for functional regulatory elements.

The standard method that has been used for phylogenetic footprinting is to construct a global multiple alignment of the orthologous regulatory sequences using a tool such as CLUSTAL W (Thompson *et al.*[16]), and then identify well conserved regions in the alignment. An algorithm designed specifically for phylogenetic footprinting without resorting to global alignment has been developed by Blanchette *et al.*[17,18]

1.2 Regulatory Elements from Heterogeneous Data

As more related genomes are sequenced and our understanding of regulatory relationships among genes improves, we will find ourselves in a situation with richer data sources than in the past. Namely, the data to be analyzed will often be *heterogeneous*, a collection of coregulated genes from one genome together with their orthologous genes in several related genomes. There is an obvious advantage to considering heterogeneous

data when it is available: namely, motifs may not be detectable when one considers only the coregulated regions from one genome or only the orthologous regions of one gene (McGuire *et al.*[19], Wang and Stormo[20]).

The most obvious way to handle heterogenous data is to treat all the regulatory regions identically: pool all the input sequences, and search for overrepresented motifs. This is precisely what was done in studies by Gelfand *et al.*[21] and McGuire *et al.*[19] There are several reasons why treating the heterogenous data homogeneously in this way discards valuable information that may be necessary for accurate prediction of regulatory elements:

1. This method ignores the phylogeny underlying the data so that, for example, similar sequences from a subset of closely related species will have an unduly high weight in the choice of motifs predicted.

2. Phylogenetic studies such as that of Lane *et al.*[22] show that instances of orthologous regulatory elements, because they evolved from a common ancestral sequence, tend to be better conserved than instances across coregulated genes of the same genome. By pooling all the sequences, this distinction is lost.

3. Perhaps most importantly, the number of occurrences of a given regulatory element will vary greatly across putatively coregulated genes: some regulatory regions will contain no occurrences, while others will contain multiple occurrences. This variance in number should be much less across orthologous genes, again because they are evolved from a single ancestral sequence. By pooling all the sequences, this distinction too is lost.

Another method for exploiting heterogeneous data involves two separate passes. For instance, Wasserman *et al.*[23], Kellis *et al.*[24], Cliften *et al.*[25], and Wang and Stormo[20] search for well conserved motifs across the orthologous genes and then, among these, search for overrepresented motifs. GuhaThakurta *et al.*[26] do the opposite, searching for overrepresented motifs in one species and eliminating those that are not well conserved in the orthologs. In both cases, the first pass acts as a filter before performing the second pass, and a drawback is that the true motif may be filtered out because it is not conserved well enough in the dimension of the first pass. In other words, these algorithms do not integrate all the available information from the very beginning.

In this paper we propose the first algorithm that uses the heterogeneous sequence data in an integrated manner. We focus on the 2-species case for concreteness and efficiency, but also because of its timeliness for the study of regulons in important sequenced pairs such as human/mouse, fruitfly/mosquito, and *C.elegans*/ *C.briggsae*.

2 Expectation-Maximization for Heterogeneous Data

The Expectation-Maximization algorithm of MEME [1] is very well suited for the discovery of regulatory elements in single-species regulons. We have generalized MEME's framework and algorithm so that it is suited for the two-species heterogeneous data problem. We call the new algorithm OrthoMEME.

The inputs to OrthoMEME are sequences $X_1, X_2, \ldots, X_n$, $Y_1, Y_2, \ldots, Y_n$, where $X_1, X_2, \ldots, X_n$ are the regulatory regions of n genes from species X, and Y_i is X_i's orthologous sequence from species Y. For ease of discussion we will assume that the motif width W is fixed but, like MEME, OrthoMEME iterates over different values of W and chooses the best result. Also like MEME, OrthoMEME can be run in any of three modes: OOPS (One Occurrence Per Sequence), ZOOPS (Zero or One Occurrence Per Sequence), or TCM (zero or more occurrences per sequence). TCM mode is particularly appropriate for most regulatory element problems.

In the heterogeneous data setting, a motif occurrence in sequence i means an occurrence in X_i and an orthologous occurrence in Y_i. That is, even in TCM mode every motif occurrence consists of an orthologous pair. Accordingly, the hidden random variables are Z_{isjk}, defined to be 1 if there are orthologous motif occurrences that begin at position j of X_i and position k of Y_i, both occurrences in orientation s (either $+$ or $-$), and 0 otherwise. (An underlying assumption is that sequences outside motif occurrences are drawn from the background distributions and, in particular, are not orthologous. This is in general untrue, but for sufficiently diverged sequences the resulting inaccuracy should be minimal.)

OrthoMEME's objective is to maximize the expected log likelihood of the model, divided by the motif width, given the input sequences and hidden variables. The model parameters specify how well conserved the motif is among the sequences of species X (parameter θ, a position weight matrix), and how well conserved orthologous pairs of motif instances are (parameter η, a vector of 4×4 transition probability matrices). More specifically,

$$\theta_{jr} = \begin{cases} \Pr(\text{residue } r \text{ in background distribution}) & \text{if } j = 0 \\ \Pr(\text{residue } r \text{ at position } j \text{ of } X\text{'s occurrences}) & \text{if } 1 \le j \le W, \end{cases}$$

$\eta_{jrs} = \Pr(\text{at position } j \text{ of motif, residue } r \text{ of } X \text{ maps to residue } s \text{ of } Y)$.

There is also a corresponding parameter θ'_{0r} that specifies the background distribution in species Y. In ZOOPS and TCM modes, there is an additional parameter λ that specifies the expected frequency of motif occurrences. Let ϕ be a vector containing all the model parameters.

In classic expectation-maximization fashion, OrthoMEME alternates between E-steps (which update the expected values of the hidden variables) and M-steps (which update the model parameters). More specifically, the E-step computes $\mathrm{E}(Z_{isjk} \mid X_i, Y_i, \phi)$, where ϕ consists of the values of the model parameters computed in the previous M-step. The M-step finds the values of the model parameters ϕ that maximize the log likelihood of the model, given the input sequences and the expected values of Z_{isjk} computed in the previous E-step.

The formulas for these steps depend on the mode (OOPS, ZOOPS, TCM). For simplicity, we present only the formulas for OOPS mode. Let $X_{i,s,p}$ be the residue present at position p of strand s in sequence X_i, and let m be the length of each input sequence. Then the E-step for OOPS mode is computed as follows:

$$\mathrm{E}(Z_{isjk} \mid X_i, Y_i, \phi) = \frac{\Pr(X_i \mid Z_{isjk} = 1, \phi) \Pr(Y_i \mid X_i, Z_{isjk} = 1, \phi)}{\sum_{s,u,v} \Pr(X_i \mid Z_{isuv} = 1, \phi) \Pr(Y_i \mid X_i, Z_{isuv} = 1, \phi)},$$

where

$$\Pr(X_i \mid Z_{isjk} = 1, \phi) = \prod_{\substack{p=1 \\ p \notin \{j, \ldots, j+W-1\}}}^{m} \theta_{0 X_{i,s,p}} \prod_{p=1}^{W} \theta_{p X_{i,s,j+p-1}},$$

$$\Pr(Y_i \mid X_i, Z_{isjk} = 1, \phi) = \prod_{\substack{p=1 \\ p \notin \{k, \ldots, k+W-1\}}}^{m} \theta'_{0 Y_{i,s,p}} \prod_{p=1}^{W} \eta_{p X_{i,s,j+p-1} Y_{i,s,k+p-1}}.$$

The model parameters are evaluated in the M-step as follows. Let M_{hfg} denote the expected number of times residue f of X is mapped to residue g of Y at position h in the motif.

$$M_{hfg} = \sum_{\substack{i,s,j,k \\ X_{i,s,j+h-1}=f \\ Y_{i,s,k+h-1}=g}} \mathrm{E}(Z_{isjk} \mid X_i, Y_i, \phi),$$

$$\eta_{hfg} = \frac{M_{hfg}}{\sum_g M_{hfg}}.$$

θ is updated as in MEME.

Each E-step and M-step runs in time $O(nm^2W)$, since the number of hidden variables is $2nm^2$. This causes the algorithm to run slowly when the input sequences are long, which is an aspect of the algorithm that we are striving to improve. MEME's running time per step is $O(nmW)$.

The algorithm needs a measure to compare solutions found, in order to choose the best motif among all those found from different initial

values of ϕ and different choices of motif width W. Unlike MEME, OrthoMEME compares solutions on the basis of the expected log likelihood of the model, divided by the motif width, given the input sequences and hidden variables. That is, it uses the very evaluation function that it is optimizing. (MEME instead uses the p-value of the relative entropy of the motif instances predicted.)

There is an interesting algorithmic problem that arises only in the TCM mode of OrthoMEME and not at all in MEME. In order to produce actual motif occurrences from the final values Z'_{isjk} of $\mathrm{E}(Z_{isjk} \mid X_i, Y_i, \phi)$, OrthoMEME must choose 0 or more good orthologous pairs $(j_1, k_1), (j_2, k_2), \ldots$ for each value of i. These pairs should represent nonoverlapping occurrences whose order is conserved between the two species, that is, $j_h + W \leq j_{h+1}$ and $k_h + W \leq k_{h+1}$, for all h. For each value of i, OrthoMEME does this by retaining only those pairs (j, k) such that Z'_{isjk} exceeds a threshold, and then using dynamic programming (quite similar to that for optimal alignment) to choose those pairs that represent nonoverlapping occurrences with conserved order and maximum total value of Z'_{isjk}.

3 Experimental Results

OrthoMEME is implemented and we intend to make it publicly available. This section reports initial results of OrthoMEME on several heterogeneous data sets. All MEME and OrthoMEME motifs discussed below were among the top 3 motifs reported on those input sequences.

Tables 1–3 show the predictions of OrthoMEME on yeast regulons from *Saccharomyces cerevisiae* and their orthologs in *Saccharomyces bayanus*. The *S. cerevisiae* target genes and binding sites for these transcription factors come from SCPD [27].

The homogeneous *S. cerevisiae* data sets of Tables 1 and 2 are known to be particularly difficult: the motif discovery tools YMF [10], MEME [1], and AlignACE [5] all failed to find the known transcription factor binding sites in these *S. cerevisiae* regulons (Sinha and Tompa [28]).

Table 1 shows OrthoMEME's predictions on the genes known to be regulated by HAP2;HAP3;HAP4. There are 5 known binding sites contained in 4 target genes. MEME predicted only 1 of these binding sites (whether run on just *S. cerevisiae* sequences or on the pooled sequences of both species), whereas OrthoMEME predicted 3 using the same parameters. In this and all subsequent tables, the underlined portions of the predicted motif occurrences are the subsequences that overlap the known binding sites.

Table 2 shows OrthoMEME's predictions on the genes known to be regulated by UASCAR. There are 4 known binding sites contained in 3 target genes, all 4 of which are predicted by OrthoMEME. MEME pre-

Table 1: HAP2;HAP3;HAP4 predicted motif, OOPS mode, sequence length 600. The column labeled "Mut" shows the number of mismatches between the orthologous motif occurrences. The underlined portions of the motif occurrences are the subsequences that overlap the known binding sites. OrthoMEME missed one occurrence in each of SPR3 and CYC1. Source: SCPD [27].

		S. cerevisiae		*S. bayanus*		
Gene	Str	Pos	Instance	Pos	Instance	Mut
CYC1	+	-284	TTGGTTGG	-319	TTGGTTGG	0
SPR3	-	-485	ATGGTTGC	-377	ATGGTTGA	1
QCR8	-	-211	TTCATTGG	-225	TTTATTGG	1
COX6	-	-286	CTGATTGG	-283	CTGATTGG	0

Table 2: UASCAR predicted motif, TCM mode, sequence length 300. OrthoMEME missed no occurrences. Source: SCPD [27].

		S. cerevisiae		*S. bayanus*		
Gene	Str	Pos	Instance	Pos	Instance	Mut
CAR2	+	-218	CTCTGTTAAC	-222	CTCTGTTAAC	0
CAR2	+	-154	TGCCCTTGCC	-153	TGCCCTTGCC	0
ARG5,6	+	-114	TTCCATTAGG	-122	TTCCATTAGG	0
CAR1	+	-169	TTCACTTAGC	-176	TTCACTTAGC	0
ARG5,6	+	-52	TGCCTTTAGT	-56	TGCCTTTAGT	0
ARG5,6	+	-286	TTCACTTAAA	-294	TTCACTTAAG	1
CAR2	+	-189	TGCCGTTAGC	-193	TGCCGTTAGC	0
CAR2	-	-252	TTGCGTGTGG	-257	TTGCGTGCGG	1
ARG5,6	+	-224	ATGACTCAGT	-228	ATGACTCAGT	0
CAR1	-	-209	TGCCATTAGC	-216	TGCCGTTAGC	1
CAR1	+	-232	TGCCCTTCGC	-239	TGCCCTTGGC	1
CAR1	+	-86	TTCTCTTCTC	-73	TTCTCCTCTC	1

dicted none of these binding sites when run on the *S. cerevisiae* sequences alone, and all 4 when run on the pooled sequences of both species.

Table 3 summarizes the performance of OrthoMEME on some less difficult yeast regulons [28]. On all three regulons OrthoMEME had few true negatives. On the SCB and PDR3 regulons, OrthoMEME's number of false positives was comparable to that of MEME. On the MCB regulon, OrthoMEME had many more false positives than MEME, but many fewer true negatives to compensate.

Tables 4 and 5 give examples of OrthoMEME run on heterogeneous human/mouse data. Table 4 shows target genes of the human transcription factor SRF together with their mouse orthologs. TRANSFAC [29] reports one known binding site in each of these 4 regulatory sequences.

Table 3: Summary of other yeast regulons, *S. cerevisiae vs. S. bayanus*, TCM mode, sequence length 1000. Column headings: "genes", the number of target genes in the regulon; "known", the number of known *S. cerevisiae* binding sites in these target genes; "MEME, *S. cer.*", MEME run on the *S. cerevisiae* sequences; "MEME, pooled", MEME run on the pooled sequences of both species; "FP", the number of false positives (predictions that were not binding sites); "TN", the number of true negatives (binding sites that were not predicted). Source: SCPD [27].

			OrthoMEME		MEME, *S. cer.*		MEME, pooled	
factor	genes	known	FP	TN	FP	TN	FP	TN
SCB	3	8	6	2	8	2	13	4
MCB	5	11	10	1	5	7	6	5
PDR3	4	11	7	2	6	1	13	1

Table 4: SRF predicted motif, OOPS mode, sequence length 1000. OrthoMEME missed one occurrence in each of B-ACT and apoE. Source: TRANSFAC [29].

		H. sapiens		*M. musculus*		
Gene	Str	Pos	Instance	Pos	Instance	Mut
B-ACT	+	-73	CCTTTTATGG	-65	CCTTTTATGG	0
c-fos	-	-314	CCTAATATGG	-459	CCTAATATGG	0
apoE	-	-43	CCAATTATAG	-855	CCAATTATAG	0
CA-ACT	-	-850	CCTTATTTGG	-111	CCTTATTTGG	0

OrthoMEME predicted 2 of these 4 known binding sites. MEME, using the same parameters, found none of them, whether run on just the human sequences or on the pooled human and mouse sequences.

Table 5 shows target genes of the human transcription factor NF-κB together with their mouse orthologs. TRANSFAC [29] reports 11 known binding sites in these 10 genes. Because OrthoMEME was run in OOPS mode, it missed one of the two occurrences in IL-2. It also missed the known occurrences in SELE and IL-2Rα. MEME, using the same parameters, performed as well on this regulon.

Table 6 shows an example of OrthoMEME's predictions on a worm regulon. This is a collection of *Caenorhabditis elegans* genes regulated by the transcription factor DAF-19 (Swoboda *et al.* [30]), together with orthologs from *Caenorhabditis briggsae*. Each regulatory region in *C. elegans* is known to contain one instance of the "x-box", which is the binding site of DAF-19. OrthoMEME predicted all five of the documented x-boxes [30], as did MEME. (The full x-box has width 14 bp, of which OrthoMEME omitted the somewhat less conserved first 4 bp.)

Table 5: NF-κB predicted motif, OOPS mode, sequence length 1000. OrthoMEME missed one occurrence in each of SELE, IL-2Rα, and IL-2. Source: TRANSFAC[29].

			H. sapiens		*M. musculus*	
Gene	Str	Pos	Instance	Pos	Instance	Mut
SELE	-	-285	CCCGGGAATATCCAC	-262	TCTGGGAATATCCAC	2
ICAM-1	-	-228	CTCCGGAATTTCCAA	-250	TCTAGGAATTTCCAA	4
GRO-γ	+	-160	TCCGGGAATTTCCCT	-140	TCCGGGAATTTCCCT	0
GRO-α	+	-160	TCCGGGAATTTCCCT	-140	TCCGGGAATTTCCCT	0
IL-2Rα	-	-306	TGCGGTAATTTTTCA	-276	TGCGGTAATTTTTCA	0
GRO-β	+	-156	TCCGGGAATTTCCCT	-146	TCAGGGAATTTCCCT	1
TNF-β	+	-274	CCTGGGGGCTTCCCC	-251	CCTGGGGGCTTCCCC	0
IL-6	+	-139	TGTGGGATTTTCCCA	-125	TGTGGGATTTTCCCA	0
IFN-β	-	-140	CAGAGGAATTTCCCA	-137	CAGAGGAATTTCCCA	0
IL-2	+	-255	AGAGGGATTTCACCT	-257	AGAGGGATTTCACCT	0

Table 6: DAF-19 predicted motif, OOPS mode, sequence length 1000. OrthoMEME missed no occurrences. Source: Swoboda *et al.*[30].

			C. elegans		*C. briggsae*	
Gene	Str	Pos	Instance	Pos	Instance	Mut
che-2	+	-126	TCATGGTGAC	-178	CCATGGCAAC	3
osm-1	-	-86	CCATGGTAGC	-79	CCATGGCAAC	2
f02d8.3	-	-79	CCATGGAAAC	-93	CCATGGAAAC	0
osm-6	-	-100	CTATGGTAAC	-764	CGATGACAAA	4
daf-19	-	-109	CCATGGAAAC	-243	CTTTGGCAAA	4

4 Conclusion

As more genomes are sequenced and our understanding of regulatory relationships among genes improves, algorithms for motif discovery from the rich source of heterogeneous sequence data will become prevalent. We have introduced the first algorithm to deal with heterogeneous data sources in a truly integrated manner, using all the data from the onset of analysis.

We are still in the early stages of experimenting with the implementation and its parameters. There is much room for improved prediction accuracy and we are optimistic that, with more experience, we will consistently be able to solve problems with OrthoMEME that cannot be solved from homogeneous data alone.

There is a reasonably straightforward extension to $K > 2$ species in which the transition matrices η_j are replaced by rate matrices and one assumes that the phylogeny and its branch lengths are given. For this

extension the running time would be $O(nm^K W)$, which is prohibitive. We are working on faster algorithms for this case and also the important case $K = 2$.

For the case $K = 2$, it seems important to have a better understanding of how evolutionary distance between the species affects OrthoMEME's accuracy.

Acknowledgments

Peter Swoboda provided us with the *C. elegans* DAF-19 data set, and Phil Green and Joe Felsenstein made helpful suggestions. This material is based upon work supported in part by the Howard Hughes Medical Institute, by the National Science Foundation under grants DBI-9974498 and DBI-0218798, and by the National Institutes of Health under grant R01 HG02602.

References

1. Timothy L. Bailey and Charles Elkan. The value of prior knowledge in discovering motifs in MEME. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 21–29, Menlo Park, CA, 1995. AAAI Press.
2. Alvis Brāzma, Inge Jonassen, Jaak Vilo, and Esko Ukkonen. Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Research*, 15:1202–1215, 1998.
3. Jeremy Buhler and Martin Tompa. Finding motifs using random projections. *Journal of Computational Biology*, 9(2):225–242, 2002.
4. Gerald Z. Hertz and Gary D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7/8):563–577, July/August 1999.
5. J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, 296:1205–1214, 2000.
6. Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, and John C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 8 October 1993.
7. Charles E. Lawrence and Andrew A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Genetics*, 7:41–51, 1990.
8. Isidore Rigoutsos and Aris Floratos. Motif discovery without alignment or enumeration. In *RECOMB98: Proceedings of the Second*

Annual International Conference on Computational Molecular Biology, pages 221–227, New York, NY, March 1998.

9. Emily Rocke and Martin Tompa. An algorithm for finding novel gapped motifs in DNA sequences. In *RECOMB98: Proceedings of the Second Annual International Conference on Computational Molecular Biology*, pages 228–233, New York, NY, March 1998.
10. Saurabh Sinha and Martin Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 30(24):5549–5560, December 2002.
11. J. van Helden, A. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, 28:1808–1818, 2000.
12. C. T. Workman and G. D. Stormo. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. In *Pacific Symposium on Biocomputing*, pages 464–475, Honolulu, Hawaii, January 2000.
13. D.A. Tagle, B.F. Koop, M. Goodman, J.L. Slightom, D.L. Hess, and R.T. Jones. Embryonic ϵ and γ globin genes of a prosimian primate (*Galago crassicaudatus*) nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology*, 203:439–455, 1988.
14. Gabriela G. Loots, Ivan Ovcharenko, Lior Pachter, Inna Dubchak, and Edward M. Rubin. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Research*, 12:832–839, May 2002.
15. Dario Boffelli, Jon McAuliffe, Dmitriy Ovcharenko, Keith D. Lewis, Ivan Ovcharenko, Lior Pachter, and Edward M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, February 2003.
16. J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.
17. Mathieu Blanchette, Benno Schwikowski, and Martin Tompa. Algorithms for phylogenetic footprinting. *Journal of Computational Biology*, 9(2):211–223, 2002.
18. Mathieu Blanchette and Martin Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research*, 12(5):739–748, May 2002.
19. Abigail Manson McGuire, Jason D. Hughes, and George M. Church. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Research*, 10:744–757, 2000.
20. Ting Wang and Gary D. Stormo. Combining phylogenetic data

with coregulated genes to identify regulatory motifs. *Bioinformatics*, 2003. To appear.

21. M. S. Gelfand, E. V. Koonin, and A. A. Mironov. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Research*, 28(3):695–705, 2000.
22. Robert P. Lane, Tyler Cutforth, Janet Young, Maria Athanasiou, Cynthia Friedman, Lee Rowen, Glen Evans, Richard Axel, Leroy Hood, and Barbara J. Trask. Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proceedings of the National Academy of Science USA*, 98(13):7390–7395, June 19, 2001.
23. Wyeth W. Wasserman, Michael Palumbo, William Thompson, James W. Fickett, and Charles E. Lawrence. Human-mouse genome comparisons to locate regulatory sites. *Nature Genetics*, 26:225–228, October 2000.
24. Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423:241–254, May 2003.
25. Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A. Cohen, and Mark Johnston. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, 301:71–76, 2003.
26. Debraj GuhaThakurta, Lisanne Palomar, Gary D. Stormo, Pat Tedesco, Thomas E. Johnson, David W. Walker, Gordon Lithgow, Stuart Kim, and Christopher D. Link. Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Research*, 12:701–712, 2002.
27. Jian Zhu and Michael Q. Zhang. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, 15(7/8):563–577, July/August 1999. `http://cgsigma.cshl.org/jian/`.
28. Saurabh Sinha and Martin Tompa. Performance comparison of algorithms for finding transcription factor binding sites. In *3rd IEEE Symposium on Bioinformatics and Bioengineering*, pages 214–220. IEEE Computer Society, March 2003.
29. E. Wingender, P. Dietze, H. Karas, and R. Knüppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24(1):238–241, 1996. `http://transfac.gbf-braunschweig.de/TRANSFAC/`.
30. Peter Swoboda, Haskell T. Adler, and James H. Thomas. The RFX-type transcription factor DAF-19 regulates sensory neuron cilium formation in *C. elegans*. *Molecular Cell*, 5:411–421, March 2000.