Computational Identification and Analysis of Eukaryotic Promoters: New Algorithms on the Traces of Gene Regulation

Uwe Ohler and Heinrich Niemann

Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg

Martensstr.3, D-91058 Erlangen, Germany

eMail: Uwe.Ohler@cs.fau.de

November 13, 2000

Number of words: ca. 1850 (excluding abstract)

Number of tables: 1

keywords: DNA sequence analysis, computational promoter analysis, computational promoter finding, genome annotation, transcription regulation

software used: latex2e on linux

Abstract

With the information of the complete DNA sequence of several higher eukaryotes as well as expression patterns of thousands of genes under a variety of conditions in our hands, we now have the possibility to computationally identify and analyze the parts of a genome believed to be largely responsible for transcription control – the promoters. This article gives a short overview of the state-of-the-art techniques for promoter localization and analysis, and comments on the most recent advances in the field.

Understanding gene regulation is one of the most exciting topics within molecular genetics. To learn how the interplay among thousands of genes leads to the existance of a complex eukaryotic organism is one of the great challenges, and the availability of large amounts of information gained in the sequencing and gene expression projects both demands and enables us to use computers to solve this task.

A key role in gene regulation is played by promoter sequences. We define this here as the region proximal to the transcription start site (TSS) of protein encoding genes, those transcribed by RNA polymerase II, and leave aside distal regions such as enhancers. We want to outline the recent developments within two areas of bioinformatics that deal with promoters: The general recognition of eukaryotic promoters, and the analysis of these regions to identify the regulatory elements hidden in them. This is the first step on the way to complex models of regulatory networks. We focus on the computational point of view, pinpointing out some classic and many recent publications, and leave a more elaborate description, especially of the underlying biology, to the the cited papers and reviews.

Analyzing Promoters to Find Unknown Regulatory Elements

The interest in promoter analysis received a great boost with the arrival of microarray gene expression data: Once you have a group of genes with a similar expression profile (e.g., that are activated at the same time in the cell cycle[1]), a natural assumption is that the similar profile is (partly) caused by and reflected in a similar structure of the regulatory regions involved in transcription. The ultimate goal here is the automated construction of specific promoter models containing a combination of several regulatory elements. Research so far has focused on the detection of *single* motifs (representing transcription factor binding sites) common to the promoter sequences of putatively co-regulated genes. Although this problem might seem simple at first, it is very complex and requires that we find

- a motif of unknown size that might not be well conserved between promoters
- in a set of sequences that do not necessarily represent the complete promoters, and
- that was in many cases grouped together by a clustering algorithm that itself can be error-prone and include genes that are not co-expressed *in vivo*.

Therefore, studies have mainly concentrated on the rather "simple" genome of the budding yeast *S. cerevisiae* — it was the first fully sequenced eukaryotic organism, and the first one for which a comprehensive amount of expression data became publicly available. Statistics on mapped transcription start sites[2] show that its 5' UTR sequences are rather short (a mean of 89 bp), and most of the known regulatory elements are close to the translated part of the genes, the majority being found between 10 and 700 bp upstream from the translation start codon. This means that for yeast, the region upstream of the start codon can be used as a good approximation of a promoter region. Most algorithms searching for conserved patterns in yeast promoters thus take 500–1000 bp upstream of the start codons of supposedly co-regulated genes as data set.

There are two fundamentally different approaches to tackle the problem:

- *Alignment* methods aim at the identification of unknown signals by a significant local multiple alignment of all sequences. As a direct multiple alignment would be computationally very expensive, the methods go a different way. For example, the CONSENSUS algorithm approximates a multiple alignment by aligning sequences one by one[3] and optimizing the information content of the weight matrix constructed from the alignment. Other algorithms use a probabilistic approach; they consider the *start positions* of the motifs in the sequences to be unknown and perform a local optimization over the sequence to determine which positions deliver the most conserved motif. Two important methods are Gibbs sampling[4] and Expectation Maximization in the case of the MEME system[5].

- *Enumerative* or *exhaustive* methods examine all oligomers of a certain length and report those that occur far more often than expected from the overall promoter sequence composition[6,7]. This approach has gained in popularity since the arrival of complete genomes and is trickier than one might believe — for example, how to count patterns that overlap themselves?

From a practical point of view, the most eye-catching difference between those methods is maybe the shape of the result: The alignment approaches deliver a model of the motifs (usually a weight matrix) built from the alignment, the enumerative methods a list of over-represented oligomers, possibly already grouped to form consensus sequences. Figure 1 shows an exemplified flowchart to illustrate this.

Differences of Motif Identification Approaches

One important difference among the approaches concerns the *background* model. For example, a simple background is to account for a different overall G/C content. Without such a model, you will most likely find the obvious, e. g. mainly GC-rich motifs in organisms whose promoters have a high GC content. A better model is constructed from the set of all promoters and takes their specific sequence composition into account. With such a model, you avoid finding motifs that are common to all promoters such as TATA boxes. But this also means that a specific model, at least for each organism, has to be trained, and this information is not always available. Enumerative methods have to use such an elaborate background model because they need it to judge the importance of frequent patterns. In contrast, alignment methods usually incorporate only the G/C content, which makes them more prone to fail if the motif is not very well conserved among the sequences and the sequences to be examined become too large[8].

Most of the enumerative algorithms need to have the size of the motif specified in advance. Because of the fixed size, they often deliver a number of similar motifs simply shifted by one base or having mismatches. Some methods provide an automatic post-processing to group motifs to consensus strings and thus come up with a small number of putative regulatory elements that can be examined by experts more easily. A potential problem here is that parts of the consensus might come from different sequences.

The alignment approach requires different statistics depending on how often a pattern should be present in the sequences. For instance, MEME can be run in three modes assuming that a motif occurs exactly once, at most once, or an arbitrary number of times per promoter sequence. The Gibbs sampler implementation listed in table 1 also allows for zero or multiple ocurrences. In principle, alignment methods yield one pattern per run, but they can be run several times to detect more than one motif, masking out previously found sites. Gibbs sampling is a non-deterministic approach, meaning that even without masking out sites, it might deliver different motifs.

New directions

Some limitations of enumerative methods have been eliminated by a number of recent publications: The motif identification problem was maybe the most outstanding topic at the ISMB 2000 conference on intelligent systems for molecular biology[1] and has also been prominent in several other publications. It is now possible to detect homo- or heterodimer motifs separated by a fixed[9] or variable spacer length[10], or a variable motif length in general[11]. To allow for mismatches, ambiguous nucleotide letters (such as R for the purines) are now included in the oligomer alphabet.

Thus it seems as if the enumerative approach is the method of choice: It exhaustiveley searches over all possible oligomers and provides more significant results because of the background modeling. In practice, though, alignment methods are more flexible: Because of the simple background, they are not restricted to one specific organism. They can also find long motifs the detection of which is simply not feasible by an exhaustive approach.

[1] http://ismb2000.sdsc.edu

Also, they deliver a weight matrix as a comprehensive model for a motif which can be used more flexible than a consensus sequence for searching purposes. We therefore propose a two-step approach: First apply an enumerative approach, and use the results to initialize a weight matrix for an alignment method. Unfortunately, no such combined approach has been published yet, but the Gibbs sampler given in table 1, for example, lets you specify a weight matrix to start with. Of course, this only works if both methods are available, which so far is the case for only yeast and some microbial organisms.

The described methods are often applied on a set of promoters that were first grouped together using gene expression level measurements. Bussemaker *et al.* recently analyzed the whole set of yeast regulatory sequences without using any information on gene expression levels[11], constructing a dictionary of oligomers of increasing length and using the previous dictionary of shorter oligos as background. A new way to look at the data is to cluster genes based on *both* expression levels and common motifs at the same time[12]. This can help to separate gene groups that are active under the same conditions but belong to separate regulatory pathways.

An orthologous approach is to identify elements not by analyzing different promoters from the same organism, but promoters of the same gene from about 10 different related species[13]. An optimal alignment of a small region of specified size is constructed that takes the phylogenetic distance into account.

The question remains how we can use all these methods when we move on to the analysis of higher eukaryotes with their highly complex genomes. The euchromatin of *D. melanogaster* has a gene density of roughly one gene every 9 kilobases and an average predicted transcript size of 3058 bp[14], leaving a huge portion of the genome as potential locations of regulatory elements. In this case, the alignment of noncoding sequences from two related species, also known as phylogenetic footprinting, can help to narrow the search region and reveal conserved and potentially regulatory regions[15,16]. A recent publication closes the gap between this approach and motif identification: 28 orthologous co-regulated gene pairs from human and rat were automatically aligned to identify conserved ungapped sequence blocks, and the subsequent analysis of the conserved parts with a Gibbs sampling approach reveiled the known motifs that were missed otherwise[17]. The main assumption for phylogenetic approaches is that the regulatory pathway itself has not diverged which would result in different motifs with the same function.

If we do not have information from related species, we can concentrate on the analysis of proximal promoter regions close to TSSs, but the length of UTRs of higher eukaryotes prevents us from assuming that the TSSs can be found immediately upstream of the coding part of a gene. In our recent genome annotation assessment[18], the 92 Drosophila genes from the set for which full-length cDNA information was available had an average UTR length of about 1,900 bp (17 transcripts had UTRs longer than 1,000 bp). This means we have to find the start sites first.

Finding the Promoters in Genomic DNA

For a long time, bioinformaticians have tried to come up with algorithms able to identify the TSSs in eukaryotes. This is not easy because promoters are very diverse, and even well-known signals such as the TATA box can be weakly conserved or missing altogether. Algorithms for general promoter recognition so far can be classified into two groups:

- *Search-by-signal* algorithms make predictions based on the detection of core promoter elements such as the TATA box or the initiator, and/or transcription factor binding sites outside the core[19].

- *Search-by-content* algorithms identify regulatory regions by using measures based on the sequence composition of promoter and non-promoter (typically coding and intron sequences) examples[20].

There are also methods that combine both ideas – looking for signals and for regions of specific composition[21,22].

For an exact localization, promoter prediction should also mean identification of TSSs. But search-by-content methods do not provide good TSS predictions because they do not look for positionally conserved signals. To enable the comparison of different algorithms, predictions are thus counted as correct if they are made within a window around an experimentally verified start site. Using this scoring, an evaluation in 1997 found that many algorithms identified a third up to half of the start sites within genomic DNA sequences[23]. The programs were run *ab initio*, i. e. without any additional information but the sequence itself. The problem, though, was the vast number of false positives: Even the best algorithms had one false TSS prediction within 500–1,000 base pairs.

New Features, New Algorithms, New Hope

As a response to these rather disencouraging results, different approaches for promoter finding were pursued. One idea is to provide an accurate prediction of the TSS, but only for small regions known to contain a promoter[24]. An "opposite" algorithm provides specific predictions of regulatory regions (of a size of roughly 1,000 bp) via search-by-content, but gives no information whether the affected gene is on the leading or lagging strand, or where within the region the TSS itself is located[25]. A fundamentally different approach is to construct specific instead of general promoter models for groups of genes such as muscle active genes known by experiment to contain specific combinations of regulatory elements[26] (reviewed in detail by Werner[27]) — this is where promoter finding and analysis meet.

With the complete genome of many organisms at our fingertips, interest in general promoter prediction has awoke again. First, there is the fact that even though the *ab initio* performance of the algorithms is not as good as desired, this is not the way the annotation of genomes is done. Many algorithms are used together, and limiting the analyzed sequence to upstream from the start of a gene prediction or a cDNA alignment reduces the number of false predictions immensely. Second, slow but constant advance has been made — in a more recent assessment of both *ab initio* and gene finder coupled promoter predictions for Drosophila, the *ab initio* methods had less false positives than before[18], and coupling them with a gene finder proved to be quite successful — although it looks like it will be hard to achieve a sensitivity of more than 50%. Third, other features are known that can be derived from the DNA sequence[28] and may be suitable for promoter recognition:

- Many vertebrate promoter regions coincide with CpG islands. These are regions where the C+G content is high and the CG dinucleotide occurs more frequently than expected, a consequence of the fact that the DNA of many promoters is unmethylated so as to be accessible to regulatory proteins. A method to discriminate between CpG islands in promoters and in other parts of the genome has just been published[29]. This method can be seen as a search-by-content approach and does not deliver a TSS prediction. The use of CpG island feature in the latest version of our McPromoter predictor (see table 1) has also lead to a considerable reduction of false positives by roughly one third. Unfortunately, CpG islands only exist in vertebrate organisms.

- Features common to promoters of all organisms are structural properties of DNA, such as bendability or conformation (a compilation was carried out by Liao *et al.*[30]). For these properties, scoring tables based on di- or trinucleotides were determined experimentally and can be used to calculate profiles over the DNA sequence. Studies have shown that in general, eukaryotic promoters indeed do have a distinct profile when compared to coding or non-regulatory sequences[28]. Whether using these features will improve recognition remains to be seen: The profiles of individual sequences can be very noisy and thus not easy to use, and it is not clear yet if they provide new information that is not accurately reflected in the sequence itself.

The most recently published promoter finding and analysis tools are listed in table 1. More links can be found in comprehensive reviews[23,27].

Will we be able to find the regulatory regions of eukaryotes with high accuracy, and if so, will we be able to derive complex models for transcription regulation from their sequence? The question is open, but we certainly are on the way.

References

[1] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell-cyle regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9:3273–3297, 1998.

[2] J. Zhu and M. Q. Zhang. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, 15(7/8):607–611, 1999.

[3] G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.

[4] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[5] T. L. Baily and C. Elkan. Unsupervised learning of multiple motifs in biolpolymers using expectation maximization. *Machine Learning*, 21:51–83, 1995.

[6] J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast by computational analysis of oligonucleotide frequencies. *J Mol Biol.*, 281(5):827–842, Sep 1998.

[7] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, 8(11):1202–1215, 1998.

[8] P. Pevzner and S.-H. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. In *Proc. ISMB*, volume 8, pages 269–278, San Diego, 2000. AAAI Press.

[9] J. van Helden, A. F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, 28(8):1808–1818, 2000.

[10] S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. In *Proc. ISMB*, volume 8, pages 344–354, San Diego, 2000. AAAI Press.

[11] H. Bussemaker, H. Li, and E. D. Siggia. Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. U.S.A.*, 97:10096–10100, 2000.

[12] I. Holmes and W. J. Bruno. Finding regulatory elements using joint likelihoods for sequence and expression profile data. In *Proc. ISMB*, volume 8, pages 202–210, San Diego, 2000. AAAI Press.

[13] M. Blanchette, B. Schwikowski, and M. Tompa. An exact algorithm to identify motifs in orthologous sequences from multiple species. In *Proc. ISMB*, volume 8, pages 37–45, San Diego, 2000. AAAI Press.

[14] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson,

G. L. Gabor Miklos, H. O. Smith, E. W. Myers, G. M. Rubin, J. C. Venter, et al. The genome sequence of *Drosophila melanogaster*. *Science*, 287:2185–2195, 2000.

[15] L. Duret and P. Bucher. Searching for regulatory elements in human noncoding sequences. *Current Opinion in Structural Biology*, 7:399–406, 1997.

[16] R. C. Hardison. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.*, 16(9):369–372, 2000.

[17] W. W. Wasserman, M. Palumbo, W. Thompson, J. W. Fickett, and C. E. Lawrence. Human-mouse genome comparisons to locate regulatory sites. *Nature Gen.*, 26:225–228, 2000.

[18] M. G. Reese, G. Hartzell, N. L. Harris, U. Ohler, J. F. Abril, and S. E. Lewis. Genome annotation assessment in Drosophila melanogaster. *Genome Res.*, 10(4):483–501, 2000.

[19] D. S. Prestridge. Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol.*, 249:923–932, 1995.

[20] G. B. Hutchinson. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comp Appl Biosc.*, 12(5):391–398, 1996.

[21] V. Solovyev and A. Salamov. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. In *Proc. ISMB*, volume 5, pages 294–302, Menlo Park, 1997. AAAI Press.

[22] U. Ohler. Promoter prediction on a genomic scale — the Adh experience. *Genome Res.*, 10(4):539–542, 2000.

[23] J. W. Fickett and A. G. Hatzigeorgiou. Eukaryotic promoter recognition. *Genome Res.*, 7:861–878, 1997.

[24] M. Q. Zhang. Identifcation of human gene core promoters in silico. *Genome Res.*, 8:319–326, 1998.

[25] M. Scherf, A. Klingenhoff, and T. Werner. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol.*, 297(3):599–606, 2000.

[26] W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol.*, 278:167–181, 1998.

[27] T. Werner. Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome*, 10:168–175, 1999.

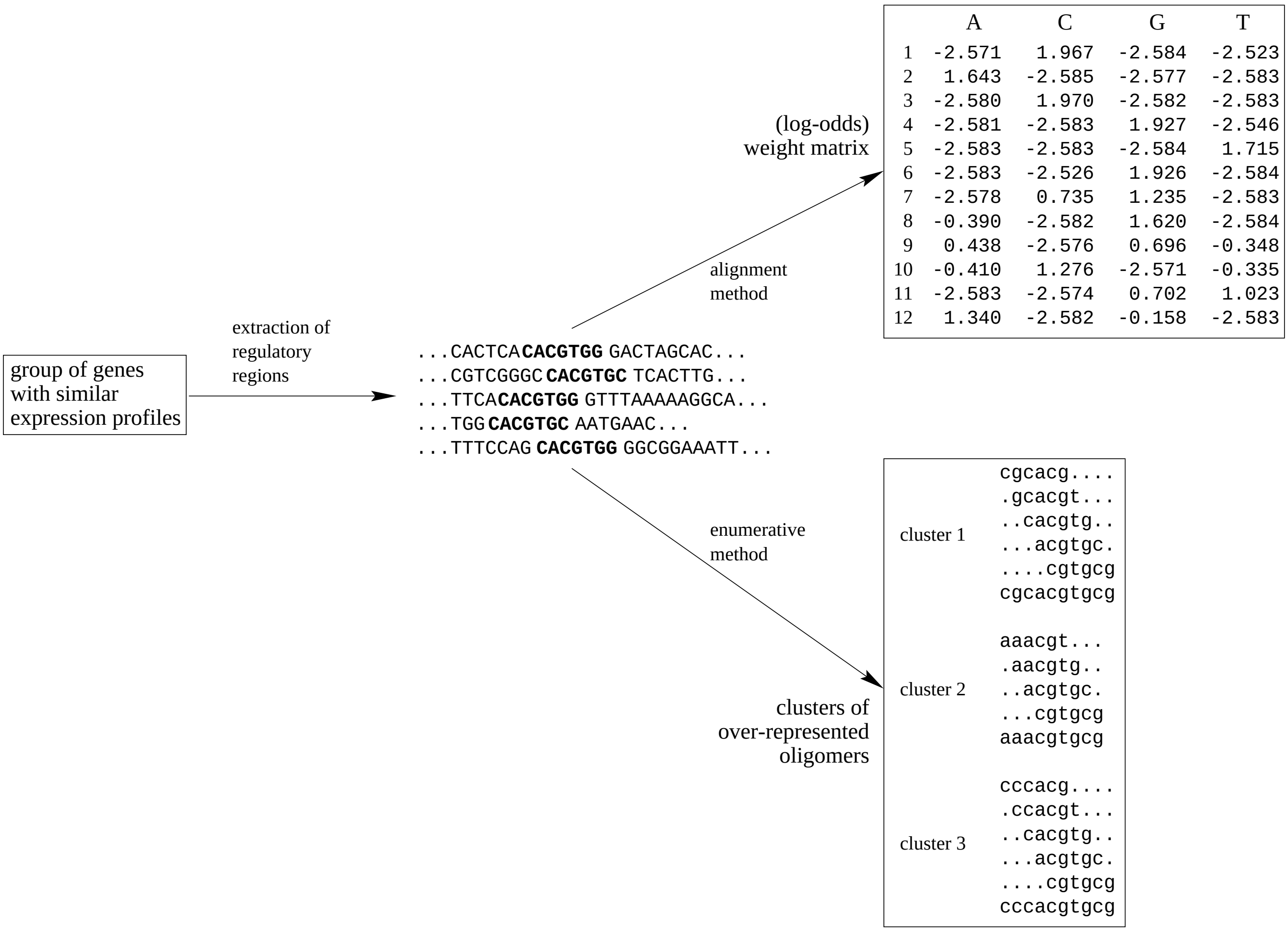
[28] A. G. Pedersen, P. Baldi, Y. Chauvin, and S. Brunak. The biology of eukaryotic promoter prediction — a review. *Comput Chem.*, 23(3/4):191–207, 1999.

[29] I. P. Ioshikhes and M. Q. Zhang. Large-scale human promoter mapping using CpG islands. *Nat. Genet.*, 26(1):61–63, 2000.

[30] G.-C. Liao, E. J. Rehm, and G. M. Rubin. Insertion site preferences of the P transposable element in *Drosophila melanogaster. Proc. Natl Acad. Sci. U.S.A.*, 97(7):3347–3351, 2000.

Program Name	**description**	**HTTP Address**
	General Promoter Finding	
Promoter2.0	search-by-signal, artificial neural network	www.cbs.dtu.dk/services/Promoter
NNPP	search-by-signal, time delay neural network	www.fruitfly.org/seq_tools/promoter.html
PromoterInspector	search-by-content, class-specific oligomers	www.gsf.de/biodv
McPromoter V3	signal/content, stochastic segment model/neural network	www.mustererkennung.de/HTML/English/ Research/Promoter
CorePromoter	signal/content, discriminant analysis	argon.cshl.org
	Promoter Analysis Tools	
RSA Tools	yeast and microbial exhaustive search	www.ucmb.ulb.ac.be/bioinformatics/rsa-tools
Gibbs sampler	alignment method	bayesweb.wadsworth.org/gibbs/gibbs.html
MEME	alignment via Expectation Maximization	meme.sdsc.edu
BBA	phylogenetic footprinting by Bayes alignment	bayesweb.wadsworth.org/ cgi-bin/bayes_align12.pl
PipMaker	phylogenetic footprinting by identity plots	bio.cse.psu.edu

Table 1: A selection of recently published WWW accessible promoter finding and analysis tools.


group of genes with similar expression profiles
extraction of regulatory regions
...CACTCA**CACGTGG**GACTAGCAC...
...CGTCGGGC**CACGTGC**TCACTTG...
...TTCA**CACGTGG**GTTTAAAAAGGCA...
...TGG**CACGTGC**AATGAAC...
...TTTCCAG**CACGTGG**GGCGGAAATT...
alignment method
(log-odds) weight matrix
enumerative method
clusters of over-represented oligomers
cluster 1
cgcacg....
.gcacgt...
..cacgtg..
...acgtgc.
....cgtgcg
cgcacgtgcg
cluster 2
aaacgt...
.aacgtg..
..acgtgc.
...cgtgcg
aaacgtgcg
cluster 3
cccacg....
.ccacgt...
..cacgtg..
...acgtgc.
....cgtgcg
cccacgtgcg

	A	C	G	T
1	-2.571	1.967	-2.584	-2.523
2	1.643	-2.585	-2.577	-2.583
3	-2.580	1.970	-2.582	-2.583
4	-2.581	-2.583	1.927	-2.546
5	-2.583	-2.583	-2.584	1.715
6	-2.583	-2.526	1.926	-2.584
7	-2.578	0.735	1.235	-2.583
8	-0.390	-2.582	1.620	-2.584
9	0.438	-2.576	0.696	-0.348
10	-0.410	1.276	-2.571	-0.335
11	-2.583	-2.574	0.702	1.023
12	1.340	-2.582	-0.158	-2.583

Figure 1: An exemplified flowchart to illustrate the two different approaches for motif identification. We analyzed 800 bp upstream from the translation start sites of the 5 genes from the yeast gene family PHO by the publicly available systems MEME (alignment) and RSA (exhaustive search, see table 1). MEME was run on both strands, one occurence per sequence mode, and found the known motif ranked as second best. RSA tools was run with oligo size 6 and non-coding regions as background, as set by the demo mode of the system.