The Analysis of Regulatory Sequences

Jacques van Helden

SCMBB. Université Libre de Bruxelles. Campus Plaine. CP 263. Boulevard du Triomphe.
B-1050 Bruxelles. Belgium. Email: jvanheld@ucmb.ulb.ac.be

CONTENTS

FOREWORDS

SCOPE OF THE COURSE

This course will deal with different aspects of the analysis of regulatory sequences.

The first lesson will consist of a general presentation of the different type of questions that can be asked about regulatory sequences, and the different approaches that can be envisaged to answer these questions (pattern-discovery, pattern matching). The second lesson will be dedicated to string-based approaches, and the third lesson to matrix-based approaches. The theoretical concepts will mainly be illustrated by concrete examples from the yeast *Saccharomyces cerevisiae*.

WEB SITE AND PRACTICAL SESSIONS

The tools developed by Jacques van Helden are available for academic users via their web interface (http://rsat.scmbb.ulb.ac.be/). During the practical session, student will apply the concepts seen during the course, and test different approaches to detect putative regulatory signals in non-coding sequences. The main resources available on the web (databases and specific sequence analysis programs) will be presented.

TRANSCRIPTIONAL REGULATION

In this chapter we briefly describe some fundamental aspects of transcriptional regulation that are relevant for the analysis of regulatory sequences. Our purpose is minimalist, and we do not pretend to review, even partially, the huge and complex field of transcriptional regulation.

THE NON-CODING GENOME

Traditionally, sequence analysis and genomics have mainly been focussed on coding sequences. These sequences however represent only a fraction of the information contained in the genome. As shown in table 1.1, the proportion of coding sequences decreases with evolution.

Organism	Year	Size	Genes	coding	non-coding
		Mb		%	%
Mycoplasma genitalium	1995	0.6	481	90	10
Haemophilus influenzae	1995	1.8	1717	86	14
Escherichia coli	1997	4.6	4289	87	13
Saccharomyces cerevisiae	1996	12	6286	72	28
Arabidiopsis thaliana	2001	120	27000	30	70
Caenorhabditis elegans	1998	97	19000	27	73
Drosophila melanogaster	2000	165	16000	15	85
Homo sapiens	2000	3000	50000	3	97

Table 1: The non-coding genome

TRANSCRIPTIONAL REGULATION

Non-coding sequences play an essential role in all cellular processes, since they mediate transcriptional regulation. Transcriptional regulation ensures the temporal and spatial specificity of expression for each gene of a genome. The level of expression is not only determined independently for each gene, but in addition, it can vary in response to a variety of signals: presence or absence of metabolites in the extra-cellular medium, inter-cellular communication, temperature, …. These signals generally provide information about the conditions outside the cell.

Transcription factors

Transcriptional regulation is mediated by classes of proteins, called *transcription factors*. These proteins interact with the general transcription machinery (RNA polymerase) in a way that either enhances (activation) or reduces (repression) the level of transcription. The same transcription factor are called dual, because they combine both effects: activate the expression of some genes while repressing the expression of other genes.

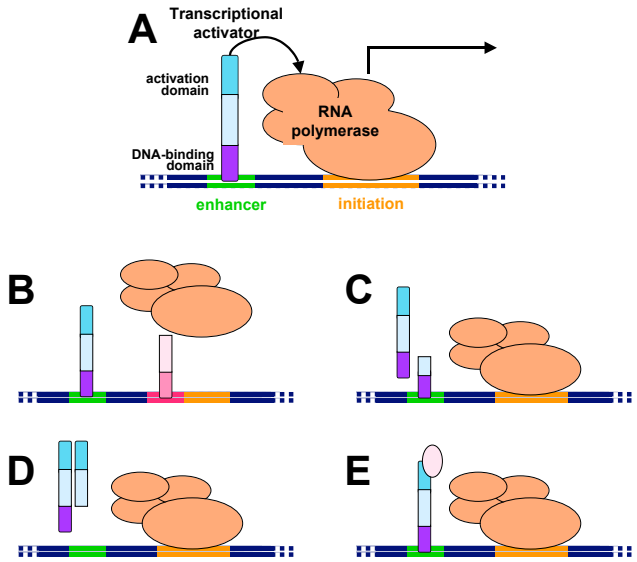


Figure 1: schematic representation of transcriptional regulation. **A:** transcriptional activation. **B:** transcriptional repression.

Transcriptional activators (Figure 1A) generally contain a domain that binds DNA in a sequence-specific manner (DNA-binding domain, and a domain that interacts with the RNA polymerase (activation domain). Repression encompasses a variety of mechanisms by which the transcription factor (repressor) reduces the expression level of a gene. Some repressors bind DNA in close vicinity (or downstream) of the transcription, and directly prevent RNA polymerase from starting transcription (Figure 1B). Another mechanism of repression is to compete with a transcriptional activator for the occupancy of the same site on DNA (Figure 1C). Some transcriptional repressors do not bind DNA at all, but rather their function is mediated by direct protein-protein interaction with a transcriptional activator. In some cases, this interaction prevents the activator from binding DNA (Figure 1D). In other cases, the repressor forms a complex with the activation domain of the transcription activator, thereby preventing its interaction with RNA polymerase (Figure 1E).

Protein-DNA interfaces

Transcription factor-DNA interfaces are generally restricted to a very few amino acids and bases. Figure 2 shows the tri-dimensional structure of some transcription factor-DNA complexes, as determined by X-ray crystallography.

Many transcription factors are active in the form of dimers, two polypeptides forming a non-covalent complex via a dimerization domain. The dimer acts on DNA like tweezers (Figure 2A,C). Each monomeric unit enters in contact with a very limited number of nucleotides (typically 3-4). In some cases, the two contact points are adjacent (Figure 2B). Several classes of transcription factors (Helix-turn-helix in bacteria, Zinc cluster proteins in fungi) contain an intermediate domain that imposes spacing between the two contact points (Figure 2D).

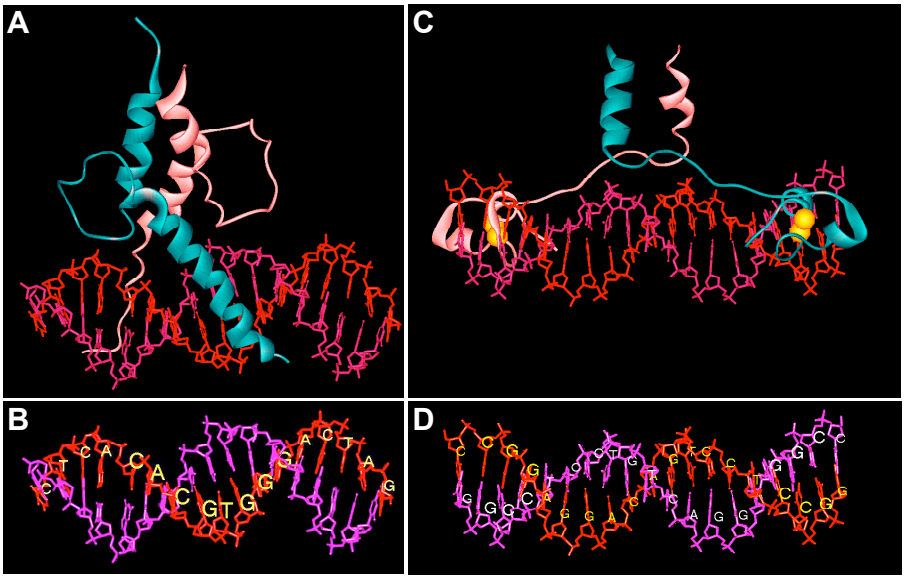


Figure 2: structure of typical transcription factor-DNA interfaces. **A:** the yeast transcription factor Pho4p forms a homodimer, which enters in contact with a set of contiguous nucleotides. **B:** sequence of nucleotides on Pho4p DNA binding site. **C:** the yeast Gal4p protein forms a homodimer, which binds a spaced pair of trinucleotides. **D:** sequence of nucleotides in the DNA binding of Gal4p.

Regulatory elements

The site on DNA where a transcription activator binds is denoted by different terms (depending on the biological field) : the yeast community favours *upstream activating sequence* (UAS), in higher organisms one speaks about *enhancers*, … The site where a repressor binds on DNA is often called *operator* (by bacteriologists), *upstream repressing sequence* (URS, in the yeast community), *silencer* (by drosophilist), … . The generic terms *cis-acting element* or *regulatory site* are used to denote the locations where transcription factors bind DNA, irrespective of their positive or negative effect on the level of expression.

Regulatory elements are very short sequences (between 5 and 30 bp) of highly conserved nucleotides. One class of regulatory element consists of a highly conserved core of 5-8 base pairs (bp), flanked by a few partly conserved bases. Another type of regulatory sites consists of a pair of very short conserved oligonucleotides (typically 3 bases) separated by a region of fixed width but variable content.

/

REPRESENTATIONS OF REGULATORY ELEMENTS

STRING-BASED REPRESENTATIONS

Different types of experiments provide primary information about the binding specificity of a DNA-binding protein. Collections of experimentally proven binding sites are stored in specialized databases such as TRANSFAC (Wingender 2004; Wingender et al. 1996), RegulonDB (Huerta et al. 1998; Salgado et al. 2001), SCPD (Zhu and Zhang 1999). These databases provide valuable information for the development and assessment of pattern detection algorithms.

Figure 3 displays a collection of binding sites for the yeast transcription factor Pho4p (Oshima et al. 1996). The table displays a qualitative estimation of the factor's binding affinity for different sequence fragments. The comparison of these sequences shows that the high affinity binding site share a "core" motif CACGTG, usually followed by a two or three cytosines (C) or guanines (G). The core CACGTG is however not sufficient to confer a high affinity: the protein does not bind to the sequences tCACGTGa or cCACGTGgaa. The lower part of the table shows two sites bound with a medium affinity, and showing a variation in the core (CACGTT) and followed by a few additional thymines (T). Despite the medium affinity, these sites have been shown to be actively involved in the regulation of the genes PHO5 and PHO84.

Gene	Site Name	Sequence	Affinity
PHO5	UASp2	---aCtCaCACACGTGGGACTAGC-	high
PHO84	Site D	---TTTCCAGCACGTGGGGCGGA--	high
PHO81	UAS	----TTATGGCACGTGCGAATAA--	high
PHO8	Proximal	GTGATCGCTGCACGTGGCCCGA---	high
PHO5	UASp3	--TAATTTGGCATGTGCGATCTC--	low
PHO84	Site C	-----ACGTCCACGTGGAACTAT--	low
PHO84	Site A	-----TTTATCACGTGACACTTTTT	low
group 1	consensus	---------gCACGTGggac-----	high-low
PHO5	UASp1	--TAAATTAGCACGTTTTCGC----	medium
PHO84	Site E	----AATACGCACGTTTTTAATCTA	medium
PHO84	Site B	-----TTACGCACGTTGGTGCTG--	low
PHO8	Distal	---TTACCCGCACGCTTAATAT---	low
group 2	consensus	--------cgCACGTTt--------	med-low
Degenerate consensus		---------GCACGTKKk-------	

Figure 3: binding sites for the Pho4p transcription factor (Oshima et al. 1996).

The collection of binding sites can be summarized with *consensus* strings such as CACGTGGG (high affinity) of CACGTTT (medium affinity). The two types of binding sites can even be represented in a more compact way, with a *degenerate consensus* CACGTKKK, where K denotes "either T of G", according to the IUPAC convention on ambiguous nucleotide code (Table 4). This representation is however an over-simplification, and suffers from several weaknesses.

1. By merging the letters G and T into the degenerate code K, we give the same weight to these letters, and we thus loose the concept that CACGTGGG is bound with a higher affinity than CACGTTT.

/

2. Several high affinity binding sites from Figure 3 do not match this consensus, and would thus be missed in a string-based search based on this pattern.
3. The degenerate consensus fails to indicate the *dependencies between successive residues*: in the collection of binding sites, the high affinity core CACGTG is usually followed by a few Gs or Cs, and the medium affinity core CACGTT by a few Ts. However, the pattern CACGTKKK would as well match sequences like CACGTGTT, CACGTGTG, CACGTTGG, which were never observed in the initial collection.

The two first limitations can be solved by using Position-Specific Scoring Matrices (PSSM), as will be shown in the next chapter. Higher-order dependencies can be treated with some more complex PSSM, or with Hidden Markov Models (HMM).

MATRIX-BASED REPRESENTATION

Position-specific scoring matrices (PSSM)

A position-specific scoring matrix (PSSM) represents the binding specificity at each position of the DNA binding site for a transcription factor. The matrix is build from an alignment of a collection of binding sites.

Each row of the matrix represents one letter of the alphabet (in this case the 4 nucleotides A, C, G and T), and each column one position of the sequence alignment. The simplest representation is a *occurrence matrix* (Table 2A), where the values in the cells indicate the absolute frequency of each residue (letter) at each position in the multiple alignment.

The weight matrix

The absolute frequency is generally not very indicative of the significance of a residue. Indeed, a general observation is most non-coding sequences are AT-rich. For instance, in the yeast *Saccharomyces cerevisiae*, the average composition of intergenic sequences is $F(A)=F(T)=0.325$, $F(C)=F(G)=0.175$. This intergenic composition can be used to estimate prior probabilities $p_A=p_T\sim0.325$; $p_C=p_G\sim0.175$.

$$W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right) \qquad f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k} \qquad \text{Equation 1}$$

Where

A	alphabet size (4 for nucleic acids, 20 for peptides)
w	matrix width (=12 in the TRANSFAC matrix $PHO4_01)
n_{ij}	occurrences of residue i in column j of the matrix
p_i	prior residue probability for residue i
f_{ij}	relative frequency of residue i at position j
k	pseudo weight (arbitrary, 1 in our example)
f'_{ij}	corrected frequency of residue i at position j

Differences in residue composition can be taken into account by calculating a *weight* ($W_{i,j}$), which represents log ratio of observed frequency ($f_{i,j}$) and prior residue probability (p_i). In addition, a *pseudo-weight (k)* can be introduced to obtain a *corrected frequency f'ij* (Hertz and Stormo 1999). The reason for introducing a pseudo-weight is that the collections of known sites used to build the matrix are generally small. For

example, the TRANSFAC matrix F$PHO4_01 (Table 2A) was calculated from no more than 8 binding sites. At some positions of the matrix, some residues have a frequency of 0 (for example the T at position 4), Using a (uncorrected) frequency of 0 would give a weight of $-\infty$, which amounts to consider as completely impossible for the factor to bind at such a position. However, the absence of this residue in our data set could either indicate that this residue hinders the factor binding, our that our current collection does not yet contain this variant for a simple reason of insufficient sampling. The introduction of the pseudo-weight resolves this problem pragmatically, since corrected frequencies cannot be null, and the weight can thus not be infinitely negative anymore. The problem is of course to estimate the importance assigned to the pseudo-weight (k) relative to the observed sites (n). A *weight matrix* (Table 2B) is derived from the occurrence matrix by calculating the weight of each residue at each position of the alignment. The weight matrix is used to assign, at each position of a sequence, a score reflecting the likelihood for the transcription factors to bind there (see chapter on *Pattern Matching*).

Information content

The information content (Hertz and Stormo 1999) is obtained by multiplying the weight by the frequency (corrected by the pseudo weight).

$$I_{i,j} = f'_{i,j} \ln\left(\frac{f'_{i,j}}{p_i}\right) \qquad I_j = \sum_{i=1}^{A} I_{i,j} \qquad I_{matrix} = \sum_{j=1}^{w}\sum_{i=1}^{A} I_{i,j} \qquad \text{Equation 2}$$

The information content can be calculated for each cell of the matrix, and then summed over rows and column to obtain I_{matrix}, the total information content of the matrix. The total information content represents the discrimination between a binding site (represented by the matrix) and the background model. Pattern discovery programs such as *consensus* (Hertz et al. 1990) select a matrix by optimizing the information content.

The information content also provides an estimate for the upper limit of the expected frequency of the binding sites in random sequences (Hertz and Stormo 1999).

$$P(site) \le e^{-I_{matrix}} \qquad \textit{Equation 3}$$

A: occurrences (counts)

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	1	3	2	0	8	0	0	0	0	0	1	2
0.175	C	2	2	3	8	0	8	0	0	0	2	0	2
0.175	G	1	2	3	0	0	0	8	0	5	4	5	2
0.325	T	4	1	0	0	0	0	0	8	3	2	2	2
1	Sum	8	8	8	8	8	8	8	8	8	8	8	8

B: frequencies

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	0.15	0.37	0.26	0.04	0.93	0.04	0.04	0.04	0.04	0.04	0.15	0.26
0.175	C	0.24	0.24	0.35	0.91	0.02	0.91	0.02	0.02	0.02	0.24	0.02	0.24
0.175	G	0.13	0.24	0.35	0.02	0.02	0.02	0.91	0.02	0.58	0.46	0.58	0.24
0.325	T	0.48	0.15	0.04	0.04	0.04	0.04	0.04	0.93	0.37	0.26	0.26	0.26
1	Sum	1	1	1	1	1	1	1	1	1	1	1	1

C: weights

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.79	0.13	-0.23	-2.20	1.05	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.175	C	0.32	0.32	0.70	1.65	-2.20	1.65	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.175	G	-0.29	0.32	0.70	-2.20	-2.20	-2.20	1.65	-2.20	1.19	0.97	1.19	0.32
0.325	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	1.05	0.13	-0.23	-0.23	-0.23
1	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.186

D: information content

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.12	0.05	-0.06	-0.08	0.97	-0.08	-0.08	-0.08	-0.08	-0.08	-0.12	-0.06
0.175	C	0.08	0.08	0.25	1.50	-0.04	1.50	-0.04	-0.04	-0.04	0.08	-0.04	0.08
0.175	G	-0.04	0.08	0.25	-0.04	-0.04	-0.04	1.50	-0.04	0.68	0.45	0.68	0.08
0.325	T	0.19	-0.12	-0.08	-0.08	-0.08	-0.08	-0.08	0.97	0.05	-0.06	-0.06	-0.06
1	Sum	0.111	0.087	0.356	1.294	0.803	1.294	1.294	0.803	0.609	0.392	0.465	0.037

Table 2: **A:** occurrence matrix representing the binding specificity of the Pho4p transcription factor from *Saccharomyces cerevisiae* (source TRANSFAC F$PHO4_01). **B:** frequencies (corrected with a pseudo-weight of 1). **C:** Weights. Positive values are shadowed. **D:** information content. Positive values are shadowed.

PATTERN DISCOVERY

INTRODUCTION

The application of pattern discovery to predict regulatory motifs can be formulated in the following way: *given a set of functionally related genes, can we detect exceptional motifs in their upstream regions, which could be responsible for their co-regulation?* This problem became very popular during the last years, due to the increasing amount of data about functional grouping of genes. A first domain of application was for the interpretation of microarray data (DeRisi et al. 1997): starting from clusters of co-expressed genes, try to predict cis-acting elements potentially responsible for their co-regulation. The same approach can be applied to other data

types such as protein complexes (Gavin et al. 2002; Ho et al. 2002), genes with similar phylogenetic profiles (Pellegrini et al. 1999), pairs of genes detected by the analysis of fusions/fission (Marcotte et al. 1999a; Marcotte et al. 1999b) (Enright et al. 1999).

The pattern discovery problem can be addressed by a variety of algorithmic approaches and statistical models. We will describe here some of these approaches, and illustrate them with selected test cases.

STUDY CASES

A simple way to evaluate a pattern discovery software is to submit a set of sequences S which contain some known motif M_{known}. The sequence is given as input for the pattern discovery program, which returns a predicted motif M_{pred}. We then compare the predicted (M_{pred}) and known (M_{known}) motifs.

As test cases, we selected the target genes of a few transcription factors from the yeast *Saccharomyces cerevisiae* (van Helden et al. 1998).

Set name	Transcription factor	Regulated genes	# genes	Description
PHO	Pho4p	PHO5; PHO8; PHO11; PHO84; PHO81	5	Activated under phosphate stress conditions
NIT	Gln3p	DAL5; GAP1; MEP1; MEP2; MEP3; PUT4; DAL80	7	Activated in response to some sources of nitrogen.
MET	Met4p	MET1; MET2; MET3; MET6; MET14; MET19; MET25; MET30; MUP3; SAM1; SAM2	11	Activated when methionine concentration is low.
GAL	Gal4p	GAL1; GAL2; GAL7; GAL80; MEL1; GCY1	6	Expressed when the yeast is fed with galactose.

Table 3 test cases for pattern discovery: list of target genes for some well-characterized transcription factors from the yeast *Saccharomyces cerevisiae*.

STRING-BASED PATTERN DISCOVERY

ANALYSIS OF WORD OCCURRENCES

We saw in the chapter 0 that the consensus of the transcription factor Pho4p consists in a short sequence of conserved residues (CACGTKKK). This is also the case for many (but not all) other transcription factors: their binding sites share a common core, consisting in a set of 5-10 contiguous residues. Starting from this observation, a simple conceptual approach to pattern discovery is to analyze the occurrences of oligonucleotides in order to detect those having an exceptionally high frequency in this input set, by comparison with some background model.

We will illustrate this approach with the test groups described above. Results obtained with some additional data sets are described in the original publication (van Helden et al. 1998).

Estimation of expected frequencies

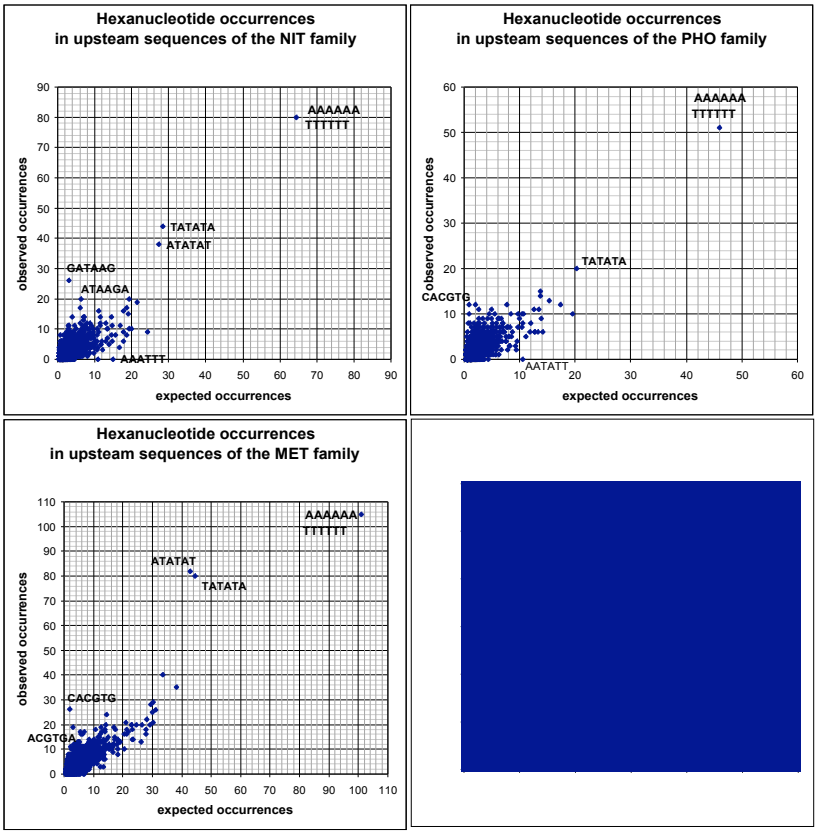
Expected occurrences were calculated on the basis of intergenic frequencies.

$$E(W)=F_{bg}(W)*T;\ T=s*(L-k+1)$$

$E(W)$	expected number of occurrences for word W
$F_{bg}(W)$	background frequency of word W. This frequency is estimated by the intergenic frequencies of the same word.
W	a given word (oligonucleotides)
k	word length (6 for hexanucleotides)
S	number of sequences in the set (5 in this case)
L	length of each sequence in the input set
T	possible positions for a k-letter word in the sequence set

Comparison of expected and observed frequencies

Figure 4 compares the expected (abscissa) and observed (ordinate) occurrences for hexanucleotides in the upstream sequences of the PHO genes.



/

plots, most words align more or less on the diagonal, with some fluctuations. The fluctuations are more important for small groups (e.g. PHO, which contains 5 genes) than for larger groups (e.g. MET, 10 genes).

On each of these plots, the most frequent pair of words is AAAAAA|TTTTTT. The next most frequent words are usually TATATA and ATATA. These words cannot be considered as over-represented, since their observed and expected occurrences are similar. This illustrates the essential difference between frequent words and over-represented words: since these frequent words are the same for all the groups, their high frequency reflects some general property of yeast upstream sequences rather than the presence of group-specific regulatory signals.

Interesting words are thus not the most frequent ones, but those which are found more frequently in the considered group than what would be expected by chance, given our background model. On the plot (Figure 4), such over-represented words appear on the top left of the diagonal. For the NIT family (Figure 4A), one pair of reverse-complementary words clearly appears as separated from the diagonal: GATAAG|CTTATC. This hexanucleotide is the so-called GATA-box, which is bound by the GATA factors, involved in nitrogen regulation. For the MET family (Figure 4C), another hexanucleotide is clearly separated from the diagonal: CACGTG, a reverse-palindrome which corresponds to the consensus of the Met4p transcription factor. For the PHO family (Figure 4B), the plot is less obvious to interpret, due to the wider overall dispersion of the cloud around the diagonal. This lower signal-to-noise ratio is due to the small number of genes in the PHO family (5 members only). However, some words seem reasonably separated from the main diagonal. In particular, CACGTG is found in 12 occurrences, whereas no more than 2 occurrences would be expected according to the background model. Consistently, this hexanucleotide corresponds to the core of the high-affinity binding sites for Pho4p. For the last group, the GAL genes, all hexanucleotides seem to align on the diagonal, suggesting that none of them is over-represented.

The graphical representation shown in Figure 4 is useful to get an intuition about the principle of word-based pattern discovery, but the simple visual comparison of observed and expected frequencies is not very accurate for selecting over-represented patterns. We saw that the hexanucleotides discarding from the diagonal correspond to regulatory signals, but where should the limit be placed?

Measuring over-representation with a P-value

We proposed a very simple probabilistic model to calculate the statistical significance of over-representation (van Helden et al. 1998).

The sequence S of length L is considered as a succession of T positions from which starts a substring of size k (word length). Since the sequence is generally linear, the number of positions for a word W of length k is smaller than L, since the last $k-1$ positions do not contain a full k-letter word.

$$T=L-k+1$$

For circular sequences (e.g. plasmids, bacterial chromosomes) the end of the string is continuous with its beginning and that a substring can be extracted from each position, so that $T=L$.

If we focus on a given word W, we can consider the sequence as a series of T trials, each of which can either result in a success (the word found at this position is W) or in a failure (the word found at this position is not W). The probability to

/

observe at least x successes (occurrences of the word W) in a succession of T trials can be calculated with the binomial probability.

$$Pvalue = P(X \geq x) = \sum_{i=x}^{T} \frac{T!}{i!(T-i)!} p^i (1-p)^{T-i}$$

Assumptions for the binomial distributions

The binomial distribution assumes that the successive trials are independent from each other and that the probability to find a word is constant over the sequence. This assumption is not properly verified, since the presence of a word of length k depends on words found at the $k-1$ preceding positions, and affects those found at the $k+1$ successive positions. For example, if the word GATAAG is found at position i of sequence S, the only words that can be found at position $i+1$ are ATAAGA, ATAAGC, ATAAGG and ATAAGT. There are thus short-term dependencies between successive words. However, when the sequence is much larger than the pattern length, and when the pattern is not self-overlapping, the hypothesis of independent positions is reasonably verified.

A notable exception to this assumption of independence is the case of self-overlapping words, like GGGGGG, TATATA, TAGTAG. Indeed, the first occurrence of such word will strongly increase the probability to find another occurrence at the following position (GGGGGG), or two (TATATA) or three (TAGTAG) positions further. This problem has been addressed by several statisticians and several corrections have been proposed. For instance, Pevzner (Pevzner et al. 1989) defined a self-overlap coefficient, which can be used to correct the estimation of variance in Gaussian models. This model relies on a normality assumption, which is verified only if the expected number of occurrences is large (>>10). In our conditions, the expectation is typically small (often smaller than 1) and Gaussian models should be avoided. Schbath (Reinert and Schbath 1998; Schbath et al. 1995) uses a compound Poisson distribution to model occurrences of clumps of words (the first occurrence being followed by overlapping occurrences of the same word).

Another way to circumvent this problem is to exclude overlapping occurrences from the counts. When the word W is found at position i of sequence S, the next occurrences of W are ignored for positions $i+1$ to $i+k-1$. The binomial schema has to be corrected accordingly: if x occurrences of word W are found, the number of possible positions for this word become

$$T=L-k+1-x(k-1) = L-(x+1)(k-1)$$

This counting mode might look like a tricky way to circumvent the problem of overlap, but it has some biological justification: the binding interface between the transcription factor and the DNA covers the whole word, and no other protein can bind simultaneously on the overlapping positions, even though the same word can be found in our string representation. We adopted this exclusion of mutually overlapping occurrences as default counting mode for web interface of the program *oligo-analysis* (van Helden 2003), but overlapping occurrences can also be counted if the user finds it appropriate according to his/her biological model.

From P-value to E-value

Another important issue is the number of words considered in a single analysis. Since the same test is simultaneously applied to all the words of the same size, the *P-*

value has to be corrected for multi-testing. The number of considered words depends on the word length, and on the counting mode (regrouping or not the pairs of reverse complements). When occurrences are counted in a strand-sensitive way, there are $D=4^k$ possible words of length k. For hexanucleotides, this makes $D=4^6=4096$ possibilities. If occurrences are counted in a strand-insensitive way, each word is regrouped with its reverse complement. For odd values of k, the number of patterns is simply divided by two: $D=4^k/2$. There are thus $4^5/2$ pairs of reverse-complementary pentanucleotides. For even values of k, the count of D is slightly more complicated. Indeed, reverse-palindromic words (e.g. CACGTG) will not be regrouped with another word. There are $4^{k/2}$ reverse-palindromes of size k (the second half of the word is determined by the first half). The total number of patterns is thus $D=(4^k-4^{k/2})/2+4^{k/2}=(4^k+4^{k/2})/2$.

A simple way to take multi-testing into account is to multiply the P-value by the number of tests (D), in order to obtain an E-value.

$$Evalue = Pvalue * D$$

The interpretation of the E-value is straightforward: it represents the expected number of false positive, given the P-value considered. For example, if we analyze hexanucleotides grouped by pairs of reverse complements and select a P-value threshold of 0.01, the *E-value* is $E=2080*0.01=20.8$, indicating that we should expect 21 false positives. This level of false positive can be easily verified by submitting random sequences to the program.

The significance score

A significance score can further be calculated from the E-value.

$$sig = -\log_{10}(Evalue)$$

This significance is convenient to interpret the over-representation: the larger is the significance, the more over-represented is the pattern. When the threshold of significance is set to 0, one expects on the average one false positive among all the words analyzed for a sequence set. With a threshold of $sig=1$, a false positive is expected every 10 sequence sets. With a threshold of $sig=s$, a false positive is expected every 10^s sequence sets.

Over-represented hexanucleotides in upstream sequences of the MET genes

Figure 5 shows the result returned by *oligo-analysis* for upstream sequences of the MET genes. Among the 2080 possible pairs of hexanucleotides, no more than 8 are statistically over-represented ($sig > 0$). The most significant word (CACGTG) corresponds to the core of the consensus for Met4p, the main regulatory of methionine metabolism in yeast. Among the 10 upstream sequences of the MET family, 9 contain at least one occurrence of this word (column *matching sequences*). In addition, some sequences contain multiple occurrences of this word, leading to a total count of 13 occurrences. The expected frequency, calculated on the whole set of yeast upstream sequences, is $F(W)=0.000164$ occurrences/positions. This word has a very high significance (sig=5.08), corresponding to a very low expected number of false positives (*E-value=8.4e-06*).

Word pair	*F(W)*	*Match. Seq.*	*occ*	*E(W)*	*P-value*	*E-value*	*sig*	*Overlaps (discarded)*	*Rank*
CACGTG|CACGTG	0.000164	9	13	1.42	4e-09	8.4e-06	5.08	0	1
CCACAG|CTGTGG	0.000265	8	11	2.30	3e-05	6.2e-02	1.21	0	2
ACGTGA|TCACGT	0.000368	9	13	3.19	3e-05	6.3e-02	1.20	6	3
AACTGT|ACAGTT	0.000610	10	17	5.28	3.8e-05	8.0e-02	1.10	0	4
ACTGTG|CACAGT	0.000374	9	12	3.24	0.00015	3.0e-01	0.52	0	5
GCTTCC|GGAAGC	0.000421	7	12	3.65	0.00042	8.6e-01	0.06	0	6
GCCACA|TGTGGC	0.000307	7	10	2.66	0.00045	9.4e-01	0.03	0	7
AGTCAT|ATGACT	0.000489	8	13	4.24	0.00046	9.6e-01	0.02	0	8

Figure 5: significant hexanucleotides in the upstream sequences of PHO genes.

The other selected words are much less significant, but we will see that another criterion suggest that they might be relevant.

Assembling words to describe more complex patterns

The 8 words selected in Figure 5 present some relationships, because some of them are mutually overlapping. For example, CACGTG can be assembled with ACGTGA to form the heptanucleotide CACGTGA. This heptanucleotide can in turn be assembled with TCACGT (the reverse complement of ACGTGA), to form the octanucleotide TCACGTGA. Among the remaining words, we also find another group of mutually overlapping words: CCACAG, CACAGT (reverse complement of ACTGTG), ACAGTT, …

The program called *pattern-assembly (van Helden 2003)* automatically assembles this type of patterns. The result of this assembly is shown in Figure 6.

```
;cluster # 1      seed: CACGTG      3 words     length
TCACGT..     ..ACGTGA     1.20
.CACGTG.     .CACGTG.     5.08
..ACGTGA     TCACGT..     1.20
TCACGTGA     TCACGTGA     5.08  best consensus

;cluster # 2      seed: CCACAG      4 words     length 8
GCCACA...    ...TGTGGC    0.03
.CCACAG..    ..CTGTGG.    1.21
..CACAGT.    .ACTGTG..    0.52
...ACAGTT    AACTGT...    1.10
GCCACAGTT    AACTGTGGC    1.21  best consensus

; Isolated patterns: 2
GCTTCC       GGAAGC       0.06
AGTCAT       ATGACT       0.02
```

Figure 6: assembly of the significant hexanucleotides selected from the MET upstream sequences.

The assembly of the 8 hexanucleotides returns two larger patterns. The first pattern (TCACGTGA, a reverse palindrome) corresponds to the binding site of Met4p, the main regulator of methionine metabolism in the yeast *Saccharomyces cerevisiae* (Thomas and Surdin-Kerjan 1997). The second pattern (GCCACAGTT|AACTGTGGC) is bound by a pair of homologous transcription factors, Met31p and Met32p, also involved in the regulation of methionine (Blaiseau et al. 1997). The two last hexanucleotides, GCTTCC and AGTCAT, cannot be

included in an assembly. Given their low level of significance, these words are likely to be false positive.

ANALYSIS OF DYAD OCCURRENCES (SPACED PAIRS OF WORDS)

A frustrating case: the GAL regulon

In this chapter, we only discussed a few examples, but the same analysis has been performed for other groups of co-regulated genes with similar results. Despite its conceptual simplicity, the program *oligo-analysis* was shown to return remarkably good results with most (but not all) yeast regulons (van Helden et al. 1998). However the analysis of oligonucleotides fails to detect the binding motif for Gal4p, and returns a negative answer: on the observed/expected frequency plot (Figure 4D), all the words align onto the diagonal. Consistently, the binomial test indicates that none of the 2080 words (grouped by pairs of reverse complement) is significantly over-represented. The failure of the program to detect the GAL-specific binding motif is particularly frustrating, since Gal4p is one of the bet characterized transcription factors in the yeast. The reason for this failure is pretty trivial: Gal4p forms a dimer, and each unit enters in contact with DNA over a few nucleotides (Figure 2C,D). The two contact points are separated by a spacing of fixed width (11bp for Gal4p), but variable content. The binding specificity is restricted to 3-4 nucleotides on each side of the spacing. One possibility would be to reduce the size of oligonucleotides, but the random expectation of trinucleotides is already quite high, so that the trinucleotides involved in the contact points of the binding sites will not be detected as significant. Another approach has been to develop a specific approach to detect over-represented pairs as a whole, as explained in the next chapter.

Analysis of spaced patterns with dyad-analysis

Spaced patterns are commonly found in transcription factor binding sites. This type of motifs are typical of some families of transcription factors, for example the fungal Zinc cluster proteins or the bacterial Helix-Turn-Helix (HTH) factors. As discussed above, word-based pattern discovery fails to detect such patterns (Figure 4D). This represents a serious inconvenient, since no less than 56 Zinc cluster proteins have been identified in the yeast genome, and in the bacteria *Escherichia coli*, most transcription factor belong to the HTH family.

In order to directly address this type of motifs, we developed a specific program, *dyad-analysis* (van Helden et al. 2000), which counts the number of occurrences of all possible spaced pairs, and compares expected and observed. Figure 7 shows the comparison of observed and expected frequencies for all pairs of trinucleotides, with all possible spacings between 0 and 16, in upstream sequences of the GAL genes. Expected frequencies were estimated as above, by counting dyad frequencies in the whole set of yeast upstream sequences (background model). As in the previous plots, most dots are more or less aligned onto the diagonal, but one dyad ($CCGn_{11}CGG$) appears clearly separated. This dyad corresponds to the two contact points of the interface between the Gal4p protein and its binding site (Figure 2).

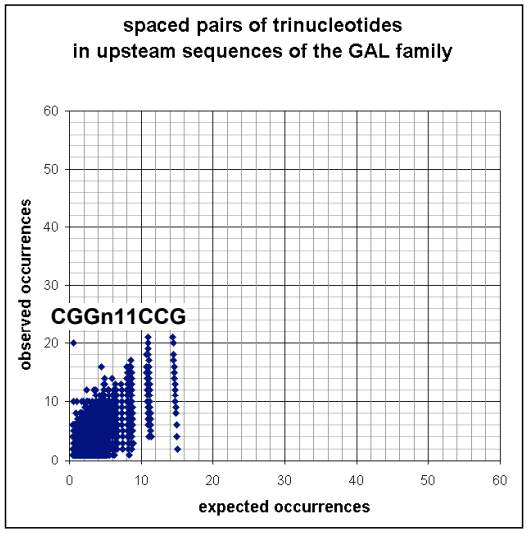


Figure 7: observed versus expected dyads in upstream sequences of the GAL genes.

We can now apply the binomial statistics as we did above for hexanucleotides. Figure 8 shows the statistically significant spaced pairs returned by the program *dyad-analysis*. In this analysis, we considered all possible pairs of trinucleotides separated by a spacing comprised between 0 and 20. In total, the number of possible dyads is $D=21*4^3*4^3=86,016$, but we regrouped them by pairs of reverse complements, so that the total number is $D=43,680$ (taking into account the number of reverse palindromes as above). Among these, no more than 6 dyads are significantly over-represented ($sig > 0$).

dyad_identifier	*F(W)*	*Occ*	*E(W)*	*P-value*	*E-value*	*sig*	*Rank*	*Ovl*
CGGn$_{11}$CCG|CGGn$_{11}$CCG	0.0000662	20	0.60	2e-12	8.9e-08	7.05	1	2
CGGn$_{12}$CGA|TCGn$_{12}$CCG	0.0000621	10	0.58	8.6e-10	3.7e-05	4.43	2	2
CGGn$_{10}$TCC|GGAn$_{10}$CCG	0.0000687	10	0.64	2.2e-09	9.8e-05	4.01	3	3
CCGn$_{01}$GCG|CGCn$_{01}$CGG	0.0000533	6	0.50	1.6e-05	6.8e-01	0.17	4	0
CCGn$_{12}$CCG|CGGn$_{12}$CGG	0.0000545	6	0.51	1.8e-05	7.7e-01	0.11	5	0
AGAn$_{05}$CCG|CGGn$_{05}$TCT	0.0001153	8	1.08	2e-05	8.8e-01	0.06	6	0

Figure 8: statistically significant dyads in upstream sequences of the GAL genes.

The most significant pattern is CGGn$_{11}$CCG|CGGn$_{11}$CCG, which appeared as the dot most distant from the diagonal in the observed/expected plot (Figure 8), and corresponds to the core of the Gal4p binding site. Several of the other selected dyads strongly overlap with this pattern. One can for example assemble CGGn$_{11}$CCG, CGGn$_{10}$TCC and CGGn$_{12}$CGA to form a larger pattern CGGn$_{12}$TCCGA. In addition, the core of the motif is reverse palindromic, and the reverse complements of the additional dyads can be included in the assembly as well (Figure 9). The resulting consensus is T**CGG**An$_8$T**CCG**A.

```
;cluster # 1     seed: CGGnnnnnnnnnnnCCG 5 words     length
;            alignt                rev_cpl     score
CCGnnnnnnnnnnnCCG.    .CGGnnnnnnnnnnnnCGG     0.11
TCGnnnnnnnnnnnCCG.    .CGGnnnnnnnnnnnnCGA     4.43
.CGGnnnnnnnnnnnCCG.   .CGGnnnnnnnnnnnnCCG.    7.05
.CGGnnnnnnnnnnnnCGA   TCGnnnnnnnnnnnnCCG.     4.43
.CGGnnnnnnnnnnTCCu..  ..GGAnnnnnnnnnnCCG.     4.01
..GGAnnnnnnnnnnCCG.   .CGGnnnnnnnnnnTCC..     4.01
TCGGAnnnnnnnnnTCCGA   TCGGAnnnnnnnnnTCCGA     7.05  best consensus

; Isolated patterns: 2
;    alignt          rev_cpl score
CCGnGCG    CGCnCGG     0.17
AGAnnnnnCCG      CGGnnnnnTCT 0.06
```

Figure 9: assembly of the statistically significant dyads detected in upstream sequences of the GAL genes.

We should keep in mind that the assembled motif is a simplification, compared to the collection of dyads. Indeed, the central dyad $CCGn_{11}CCG$ is more significant than the overlapping ones, suggesting that this might be the core of the binding interface. Searching for the complete consensus $TCGGAn_8TCCGA$ would result in the loss of some functionally active sites, because the flanking bases (T before CGG and A after it) may be present in some cis-acting elements, but absent in other ones. In order to predict the location of putative binding site, we will thus keep the collection of patterns (words or dyads) and the score associated to each of these, as illustrated in the chapter on string-based pattern matching. Besides the 3 dyads involved in the assembly, two isolated dyads are also selected. Their level of significance is however very low (0.17 and 0.06, respectively) and these are likely to be false positive.

STRENGTHS AND WEAKNESSES OF WORD- AND DYAD-BASED PATTERN DISCOVERY

Advantages

1. ***Computational efficiency.*** the computation time increases linearly with size of the input set. It can thus be applied to large sequence sets (e.g. complete genomes can be analyzed in a few minutes).
2. ***Detection of under-represented patterns.*** The same type of statistics can be applied to detect under-represented motifs, which can reveal a selective pressure for the avoidance of some functional elements. Mathias Vandenbogaert (Vandenbogaert and Makeev 2003) applied word-counting approaches to detect under-represented hexanucleotides in different bacterial genomes, and showed that the most significantly under-represented motifs correspond to restriction sites.
3. ***Exhaustivity.*** Given the relatively small number of possible solutions ($D_W=4^k$ for oligonucleotides of size k, $D_D=(s+1)*4^{2k}$ for dyads of length k with spacings between 0 and s), it is easy to calculate the P-value for each of these, and to systematically return all the over- or under-represented patterns.
4. ***Ability to return negative answers.*** The calculation of the P-value and, even better, of the E-value, allows to define significance thresholds and interpret these thresholds in terms of expected rate of false positive.

Weaknesses

1. ***Treatment of variable residues.*** A classical criticism addressed to string-based pattern discovery is that the resulting patterns (words and dyads) poorly reflect the degeneracy of the motif. In some cases (such as the PHO family above), the set of words partly reflects the degeneracy of the motif (it contains both the CACGTG and CACGTT words, as well as their surroundings). However, this is a case where the motif has two clearly distinct variants. Some motifs with a higher degree of degeneracy can be missed by the method, because none of the possible variants is significant alone.
2. ***Pattern matching.*** Pattern discovery is generally followed by pattern matching, i.e. trying to identify the positions of the discovered patterns, in order to predict putative regulatory elements. It is easy to detect the positions of the significant words and dyads obtained by the above methods, but most of their occurrences will not really correspond to motifs. Indeed, each word or dyad generally reflects only a fragment of the motif, but it is also expected to occur in other places of the sequence.

STRING-BASED PATTERN MATCHING

A simple string-based pattern matching generally gives poor predictions for transcription binding sites, for the obvious reasons that a single string-based representations fails to capture the probabilistic aspect of binding site variability, as discussed above.

The results can however be improved by matching a collection of mutually overlapping patterns (word or regular expressions), instead of a single regular expression. Multiple patterns can be used to represent overlapping fragments of a larger binding site, or the variants arising from the degeneracy of the consensus. Collections of mutually overlapping patterns can also be used to match complex motifs with higher order dependencies between neighbouring positions. For example, the following combination of words: CACGTG, ACGTGG, CGTGGG, CACGTT and ACGTTT, would capture the two variants of Pho4p binding sites (CACGTGGG and CACGTGTTT), but not the mixtures of G and T after the binding core. Such collections of mutually overlapping words are typically detected with string-based pattern discovery approaches, as we will see below. The matching can also be improved by assigning a weight to different patterns of a collection. This allows one to distinguish the strongly constrained core of the binding site (e.g. CACGTG, CACGTT) from the flanking positions, which are more degenerated (CACGTGgg, CACGTTtt). The result of such a search can be represented graphically on a feature-map (Figure 10). Annotated binding sites (green horizontal boxes) are generally denoted by a clump of mutually overlapping hexanucleotides belonging to the collection of predicted patterns.

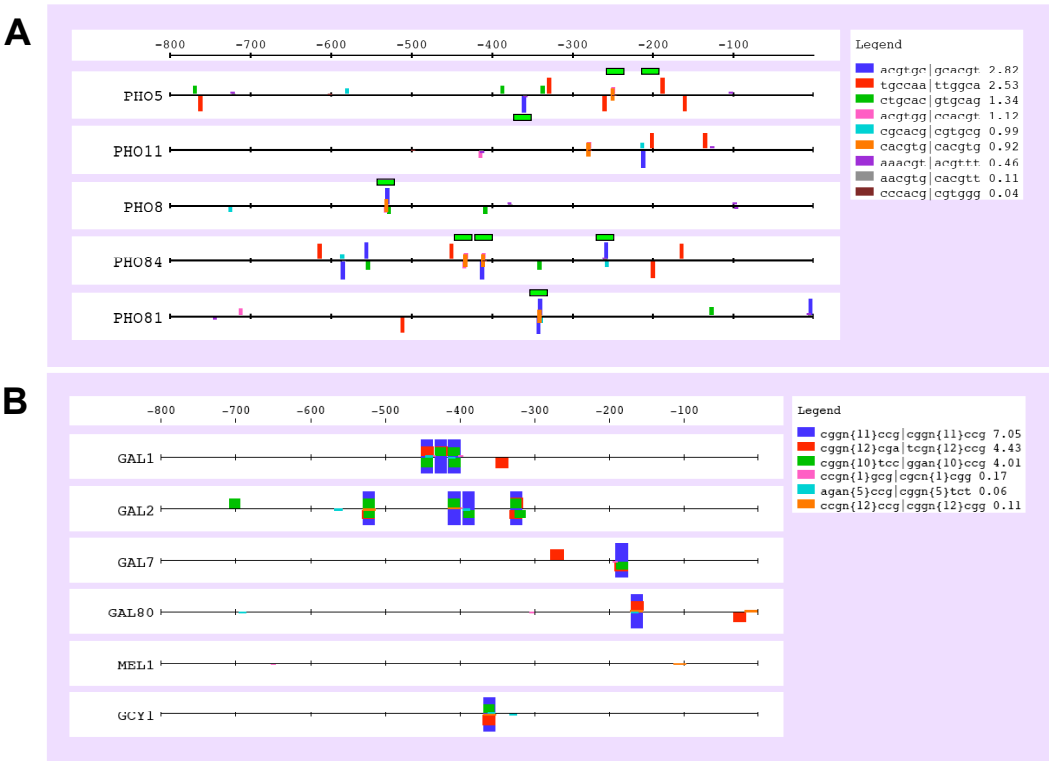


Figure 10: feature-map of pattern matching with a collection of words (A) and dyads (B). A specific weight was assigned to each pattern according to its significance in pattern discovery. **A:** over-represented hexanucleotides in upstream sequences of the PHO genes. The wider grey boxes above and below the maps indicate experimentally proven binding sites for the factor Pho4p. **B:** over-represented dyads in upstream sequences of the GAL genes.

Another possible refinement of string-based pattern matching is to allow a certain number of substitutions (mismatches). This possibility is however generally not recommended, since it would consider as equivalent any substitution at any position of the pattern. This does not correspond to the typical DNA-protein interfaces, which impose some strong constraints on specific positions, whereas other positions may show some variability. This type of position-specific variability is typically treated by matrix-based pattern matching.

MATRIX-BASED PATTERN DISCOVERY

Let us consider a simple case: we want to build a matrix of width $w=10$ with $n=12$ sequences of length $L=1000$ each. The number of possible solutions to this very small-sized test case can be estimated easily.

A first option would be to consider that each sequence should contain exactly 1 site, on either of both strands (direct or reverse). From each sequence, we need to select one among the $T=2*(L-w+1)=1,982$ possible positions for a substring of size 10. The number of possible matrices is $D=T^n=1,982^{12}=3.67e+39$.

Another option would be to consider that some sequences can contain several sites, whereas other might not contain a single site. In this case, the 12 sites can be chosen within the whole set of sequences, representing T=2n(L-w+1)=23,784 possible positions. The number of possible matrices is $C_T^n = C_{23,784}^{12} = 6.82\text{e}+43$.

/

This estimation illustrates a fundamental difficulty of matrix-based pattern discovery: the number of PSSM which could be made, even from a small sequence set, raises astronomical numbers, so that it is impossible to analyze them all in order to select the most significant one. Consequently, all the matrix-based pattern discovery programs are intrinsically condemned to scan a subset of possibilities, and return the best possible solution among this subset. The "goodness" of a matrix is generally estimated by a score (typically the information content). Various strategies have been developed to optimize the information content of a matrix extracted from a sequence set. In this course, we will present two of these strategies: a greedy algorithm developed by Hertz and Stormo (Hertz et al. 1990; Hertz and Stormo 1999; Stormo and Hartzell 1989), and a gibbs sampling algorithm originally developed by Newald and Lawrence (Lawrence et al. 1993; Neuwald et al. 1995; Neuwald et al. 1997).

CONSENSUS: A GREEDY APPROACH

A greedy algorithm has been implemented by Jerry Hertz (Hertz et al. 1990; Hertz and Stormo 1999; Stormo and Hartzell 1989) in a program named *consensus*. The principle is to start the matrix with two sequences only, and to incorporate the other sequences one by one. At each step, a subset of matrices with the highest information content are retained for the next iteration.

If the sequences have a length of, say $L=1000$ and a matrix of width $w=10$, there are $T=L-w+1=991$ possible sites in each sequence, and thus $T^2= 982,081$ possible matrices made of one site from the first sequence and one site from the second sequence.

```
MATRIX 1
number of sequences = 5
unadjusted information = 12.264
sample size adjusted information = 28.1942
ln(p-value) = -40.0503   p-value = 4.03996E-18
ln(expected frequency) = -3.91122   expected frequency = 0.0200161
A |   1   2   0   5   0   0   0   0   0   0
C |   3   0   5   0   5   0   0   0   1   2
G |   0   3   0   0   0   5   0   5   4   3
T |   1   0   0   0   0   0   5   0   0   0
  1|1   :   1/546   CACACGTGGG
  2|2   :   2/516   CACACGTGGG
  3|5   :  -3/265   TGCACGTGGC
  4|3   :   4/385   AGCACGTGGG
  5|4   :  -5/455   CGCACGTGCC

MATRIX 2
number of sequences = 5
unadjusted information = 12.2136
sample size adjusted information = 28.1438
ln(p-value) = -39.6863   p-value = 5.81381E-18
ln(expected frequency) = -3.54722   expected frequency = 0.0288047
A |   0   2   0   5   0   0   0   0   0   0
C |   4   0   5   0   5   0   0   0   1   2
G |   0   3   0   0   0   5   0   4   4   3
T |   1   0   0   0   0   0   5   1   0   0
  1|1   :   1/546   CACACGTGGG
  2|2   :   2/516   CACACGTGGG
  3|5   :  -3/265   TGCACGTGGC
  4|3   :   4/212   CGCACGTTGG
  5|4   :  -5/455   CGCACGTGCC

MATRIX 3
number of sequences = 5
unadjusted information = 12.0546
sample size adjusted information = 27.9848
ln(p-value) = -38.5478   p-value = 1.8151E-17
ln(expected frequency) = -2.40873   expected frequency = 0.0899295
A |   1   2   0   5   0   0   0   0   0   0
C |   2   0   5   0   5   0   0   0   1   1
G |   1   3   0   0   0   5   0   5   4   4
T |   1   0   0   0   0   0   5   0   0   0
  1|1   :   1/546   CACACGTGGG
  2|2   :   2/516   CACACGTGGG
  3|5   :  -3/265   TGCACGTGGC
  4|3   :   4/385   AGCACGTGGG
  5|4   :   5/455   GGCACGTGCG
```

Figure 11: the 3 matrices with the highest information content detected by the program consensus in upstream sequences of the PHO genes.

Figure 11 illustrates the result returned by the program *consensus* with upstream sequences of the PHO genes. The three top motifs are actually very similar to each other, and they match the high-affinity binding site of Pho4p (CACGTGGG). The program failed to detect medium affinity variants (CACGTTtt). An important feature of *consensus* is that a P-value and an E-value (expected frequency) are calculated for each matrix. The E-value is very informative, since it corrects the P-

value for multi-testing (as discussed above), by taking into account the number of matrices analyzed. The E-value indicates the number of false positives expected for a given P-value. For the top motif (described under MATRIX 1 in Figure 11), the P-value is very low (4.03e-18) but the E-value is 0.02 indicating that such a level of significance would be expected 2 times out of 100 random analyses. In this case, the E-value is still low, and the motif can be considered as significant.

GIBBS SAMPLING

The *gibbs* program was initially developed to discover motifs in sets of unaligned protein sequences (Lawrence et al. 1993; Neuwald et al. 1995; Neuwald et al. 1997). In short, the *gibbs* sampler is a stochastic version of the Expectation-Maximization (EM) algorithm. To initialize the program, a PSSM is built from a set of random sites collected from the input sequence. At this stage, the matrix is thus not expected to contain any specific information. After this initialization, the program iterates between a *sampling* step and a *predictive update*. During the *sampling* step, a score is assigned to each position of the input set. A random position is selected at random, with probabilities proportional to the score. During the *predictive update* step, the selected site is integrated in the matrix, from which another site is removed.

Since the initial positions were chosen at random, the initial matrix is not supposed to contain any information. During the subsequent sampling step, the scores are thus not very informative, and the selection of the next site is mainly random. During a certain number of iterations, the information content of the matrix remains thus quite low. However, if, by chance, an occurrence of the motif is incorporated in the matrix during a sampling step, it will slightly bias the next sampling step in favour of other occurrences of the same motif. And if, due to this slight bias, a second occurrence is incorporated, the bias will be reinforced. The sampler thus tends to incorporate a third, then a fourth, … occurrence of the motif, and the sampler rapidly converges towards a PSSM with high information content.

Although the original gibbs sampler (Lawrence et al. 1993; Neuwald et al. 1995; Neuwald et al. 1997) was already able to analyze DNA sequences, it had not been optimized for this task. Given the remarkable results obtained with this approach on proteins and DNA sequences, several other groups implemented their own version of a DNA-dedicated *gibbs* sampler, with various improvements:

1. Possibility to search patterns on boths strands.
2. Possibility to search multiple motifs, with iterative masking (sites used in a motif cannot be re-used for a subsequent motif).
3. Calculation of additional scores (information content, MAP, …)
4. Background models based on Markov chains of arbitrary order.

```
#INCLUSive Motif Model v1.0
#
#ID = box_1_1_ACGTGCnnmn
#Score = 41.4
#W = 10
#Consensus = ACGTGCnnmn
0.980384        0.0053098       0.00515859      0.00914785
0.00933481      0.976359        0.00515859      0.00914785
0.00933481      0.0053098       0.976208        0.00914785
0.138808        0.134783        0.00515859      0.72125
0.00933481      0.0053098       0.976208        0.00914785
0.00933481      0.717412        0.264105        0.00914785
0.268281        0.0053098       0.523051        0.203358
0.527228        0.0053098       0.328842        0.138621
0.591964        0.393729        0.00515859      0.00914785
0.203545        0.0053098       0.199368        0.591777

#ID = box_1_2_CsCACGTknk
#Score = 28.7803
#W = 10
#Consensus = CsCACGTknk
0.205241        0.773606        0.00762748      0.013526
0.109522        0.390728        0.390505        0.109245
0.0138024       0.965044        0.00762748      0.013526
0.970995        0.00785106      0.00762748      0.013526
0.0138024       0.965044        0.00762748      0.013526
0.0138024       0.00785106      0.964821        0.013526
0.0138024       0.00785106      0.00762748      0.970719
0.0138024       0.00785106      0.486224        0.492123
0.0138024       0.19929         0.199066        0.587842
0.0138024       0.19929         0.390505        0.396403

#ID = box_1_3_GCTGnTnTTs
#Score = 9.30447
#W = 10
#Consensus = GCTGnTnTTs
0.0152634       0.0086821       0.961097        0.0149577
0.121115        0.855493        0.00843485      0.0149577
0.0152634       0.0086821       0.00843485      0.96762
0.0152634       0.0086821       0.961097        0.0149577
0.332817        0.537939        0.00843485      0.120809
0.0152634       0.0086821       0.00843485      0.96762
0.226966        0.114533        0.643543        0.0149577
0.121115        0.0086821       0.00843485      0.861768
0.0152634       0.0086821       0.00843485      0.96762
0.0152634       0.432087        0.43184         0.120809
```

Figure 12: 3 top motifs discovered in upstream sequences of the PHO genes with MotifSampler. The Markov order of order 5 was generated with all the yeast upstream sequences. The program was used with the following options:
MotifSampler –f PHO_up800.fasta –b mkv5_yeast_allup800_noorf.txt –s 1 –n 3 –w 10 –x 1 –r 1

Figure 12 illustrates the result obtained with Gert Thijs' MotifSampler (Thijs et al. 2001) on upstream sequences of the PHO genes. For this analysis, we used a Markov chain of order 5. Actually, this is equivalent to a calibration of expected frequencies based on hexanucleotides frequencies. Motifs were searched on both strands, with a width of 10 bp. For each motif, the program returns the consensus,

/

followed by a frequency matrix (the frequency matrix is presented vertically: rows correspond to positions, columns to residues). The top motif (consensus ACGTGCnnmn) matches the PHO4p consensus (CACGTKkk), but it is shifted rightwards, so that the beginning of the motif is missing. The second motif (CsCACGTknk) has a weaker score, but it is better centred, and it reflects the degeneracy of the right side of the Pho4p consensus (CACGTG or CACGTK).

STRENGTHS AND WEAKNESSES OF MATRIX-BASED PATTERN DISCOVERY

Matrix-based pattern discovery presents the advantage of returning a probabilistic description of motif degeneracy: the matrix indicates the frequency of each residue at each position of the motif. The main difficulty is in the choice of appropriate parameters: most programs require for the user to specify the matrix width, and the expected number of site occurrences. Since this information is typically not provided, the user has to make guesses, or to try various possibilities and select the most convincing result.

The greedy approach, implemented in the program *consensus,* returns good results (at least with microbial data sets used in our tests), but is sensitive to the order of the sequences in the input set. If, for some reason, the first sequence does not contain any occurrence of the motif, the program will not be able to recover it subsequently.

One advantage of the gibbs sampler is time efficiency: large sequence sets can be treated in a few seconds. In comparison to the EM algorithm, the gibbs sampler shows a better ability to avoid suboptimal solutions (local optima), due to the stochastic sampling. A drawback of this is that independent runs of the program are expected return different motifs, even if the same input sequence has been analyzed with the same parameters. The program can easily be stuck in suboptimal solutions, like AT-rich motifs. The choice of a higher order Markov model is essential to reduce this effect.

CONCLUDING REMARKS

The aim of this chapter was to give a short introduction to the prediction of regulatory signals in non-coding sequences. This introduction is incomplete and biased. Incomplete because a whole book would be necessary to describe the multitude of approaches developed to detect motifs in biological sequences. Biased because I deliberately placed a stronger emphasis on string-based pattern discovery approaches, firstly because these are conceptually simpler and secondly because, as developer of two of them, I know them better.

Since a few years, the decryption of regulatory signals has been recognized as a major challenge to interpret genome information, and many researchers have joined the field. Besides the methodological issues (which algorithm should be chosen, with which parameters, etc.), the availability of an increasing number of genomes has opened the door to a perspective which was out of reach no more than 5 years ago: applying comparative genomics to understand the evolution of gene regulation. This perspective is particularly exciting for higher organisms, since morphological differences are probably to be found in gene regulation rather than in protein structures themselves. But we are far from there: if some pattern discovery methods return decent results with sets of co-regulated genes from microbial organisms, the

detection of signals in mammalian genomes is still in its infancy, and the rates of false positives are currently so high that the results are barely interpretable. There is no doubt that the future will be paved of exciting developments and discoveries for bioinformaticians willing to face this challenge.

PRACTICAL SESSIONS

A series of tutorials and exercises can be found at http://rsat.scmbb.ulb.ac.be/rsat/.

ANNEXES

IUPAC AMBIGUOUS NUCLEOTIDE CODE

A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A or G	puRine
Y	C or T	pYrimidine
W	A or T	Weak hydrogen bonding
S	G or C	Strong hydrogen bonding
M	A or C	aMino group at common position
K	G or T	Keto group at common position
H	A, C or T	not G
B	G, C or T	not A
V	G, A, C	not T
D	G, A or T	not C
N	G, A, C or T	aNy

Table 4 IUPAC Ambiguous nucleotide code

REFERENCES

Blaiseau, P.L., A.D. Isnard, Y. Surdin-Kerjan, and D. Thomas. 1997. Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol Cell Biol* 17: 3640-3648.

DeRisi, J.L., V.R. Iyer, and P.O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680-686.

Enright, A.J., I. Iliopoulos, N.C. Kyrpides, and C.A. Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402: 86-90.

Gavin, A.C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141-147.

Hertz, G.Z., G.W.d. Hartzell, and G.D. Stormo. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci* 6: 81-92.

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

Ho, Y., A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sorensen, J. Matthiesen, R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W. Hogue, D. Figeys, and M. Tyers. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415: 180-183.

Huerta, A.M., H. Salgado, D. Thieffry, and J. Collado-Vides. 1998. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res* 26: 55-59.

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

Marcotte, E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. 1999a. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285: 751-753.

Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999b. A combined algorithm for genome-wide prediction of protein function. *Nature* 402: 83-86.

Neuwald, A.F., J.S. Liu, and C.E. Lawrence. 1995. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* 4: 1618-1632.

Neuwald, A.F., J.S. Liu, D.J. Lipman, and C.E. Lawrence. 1997. Extracting protein alignment models from the sequence database. *Nucleic Acids Res* 25: 1665-1677.

Oshima, Y., N. Ogawa, and S. Harashima. 1996. Regulation of phosphatase synthesis in Saccharomyces cerevisiae--a review. *Gene* 179: 171-177.

Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* 96: 4285-4288.

Pevzner, P.A., M. Borodovsky, and A.A. Mironov. 1989. Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics

and prediction of the frequencies of occurrence of words. *J Biomol Struct Dyn* 6: 1013-1026.

Reinert, G. and S. Schbath. 1998. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J Comput Biol* 5: 223-253.

Salgado, H., A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez, and J. Collado-Vides. 2001. RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res* 29: 72-74.

Schbath, S., B. Prum, and E. de Turckheim. 1995. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J Comput Biol* 2: 417-437.

Stormo, G.D. and G.W.d. Hartzell. 1989. Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci U S A* 86: 1183-1187.

Thijs, G., M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17: 1113-1122.

Thomas, D. and Y. Surdin-Kerjan. 1997. Metabolism of sulfur amino acids in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* 61: 503-532.

van Helden, J. 2003. Regulatory sequence analysis tools. *Nucleic Acids Res* 31: 3593-3596.

van Helden, J., B. Andre, and J. Collado-Vides. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 281: 827-842.

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

Vandenbogaert, M. and V. Makeev. 2003. Analysis of bacterial RM-systems through genome-scale analysis and related taxonomy issues. *In Silico Biol* 3: 127-143.

Wingender, E. 2004. TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol* 4: 55-61.

Wingender, E., P. Dietze, H. Karas, and R. Knuppel. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 24: 238-241.

Zhu, J. and M.Q. Zhang. 1999. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* 15: 607-611.