Regulatory Sequence Analysis

Matrix-based approaches for pattern discovery

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Alignment of transcription factor binding sites

Binding sites for the yeast Pho4p transcription factor

(Source : Oshima et al. Gene 179, 1996; 171-177)

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-------	

Regulatory sites : matrix description

Alignment matrix

Pos Base	1	2	3	4	5	6	7	8	9	10	11	12
A	1	3	2	0	8	0	0	0	0	0	1	2
C	2	2	3	8	0	8	0	0	0	2	0	2
G	1	2	3	0	0	0	8	0	5	4	5	2
T	4	1	0	0	0	0	0	8	3	2	2	2
			V	C	A	C	G	T	K	B		

Binding site for the yeast Pho4p transcription factor

(Source : Transfac matrix F$PHO4_01)

Position-weight matrix

Prior	Pos	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00
0.33	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
0.18	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
0.18	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
0.33	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	0.19

$$W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}$$

A	*alphabet size (=4)*
p_i	*prior residue probability for residue i*
$f_{i,j}$	*relative frequency of residue i at position j*
k	*pseudo weight (arbitrary, 1 in this case)*
$f'_{i,j}$	*corrected frequency of residue i at position j*

Reference: Hertz (1999). Bioinformatics 15:563-577.

Information content

Prior	Pos	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00
0.33	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
0.18	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
0.18	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
0.33	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	0.11	0.09	**0.36**	**1.29**	**0.80**	**1.29**	**1.29**	**0.80**	**0.61**	**0.39**	**0.47**	0.04

$$I_{matrix} = \sum_{j=1}^{w} \sum_{i=1}^{A} I_{i,j}$$

$$I_{i,j} = f'_{i,j} \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}$$

A	*alphabet size (=4)*
w	*matrix width (=12)*
p_i	*prior residue probability for residue i*
$f_{i,j}$	*relative frequency of residue i at position j*
k	*pseudo weight (arbitrary, 1 in this case)*
$f'_{i,j}$	*corrected frequency of residue i at position j*

Reference: Hertz (1999). Bioinformatics 15:563-577.

Pattern discovery: typical dimensionality

- Typical case: GAL genes
  - s — 6 genes
  - occ_e — expected pattern occurrence: 12
  - L — 800 bp upstream sequences
analysis on both strands
  - w — matrix width = 25
- Let us assume that
  - A signal can be found on any strand
  - Each sequence contains one or several occurrences

$$N_{alignments} = C_{2s(L-w+1)}^{occ_e} = 8.8 * 10^{38}$$

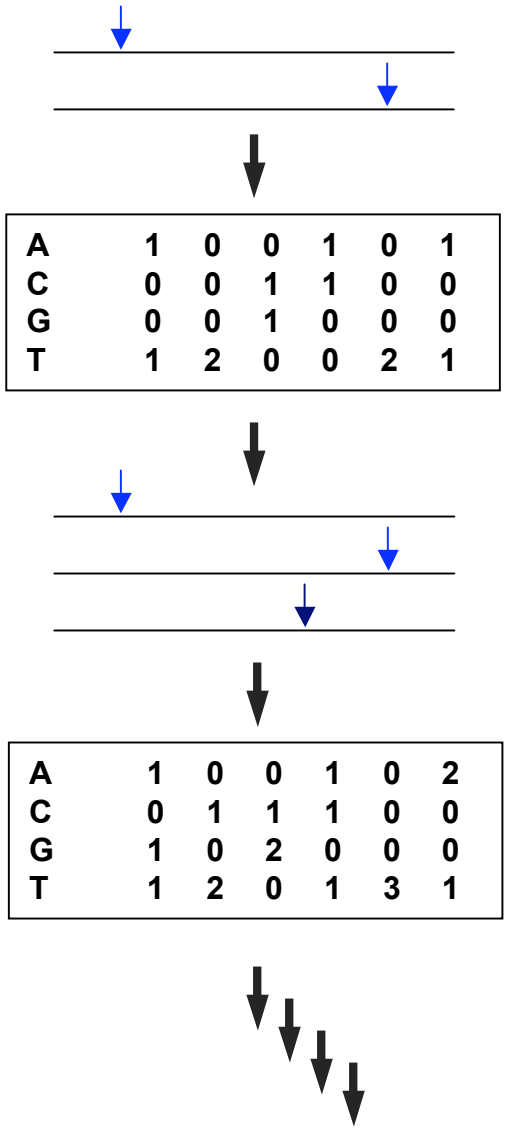
Matrix-based pattern discovery

- Problem: the number of possible matrices is too large to be tractable
- Approaches: define heuristics to extract a matrix with highest possible information content (lowest probability to be due to random effect) → optimization techniques
- Two approaches working with regulatory sequences
  - greedy algorithm
  - gibbs sampling

Pattern discovery: greedy algorithm

(consensus, by Jerry Hertz)

1) Create all possible matrices with two sequences
2) Retain the most significant matrices only
3) Find best match in next sequence and incorporate it into the matrix
4) Iterate from (2) untill all sequences are incorporated
5) Return the most significant matrices



Greedy algorithm: weaknesses

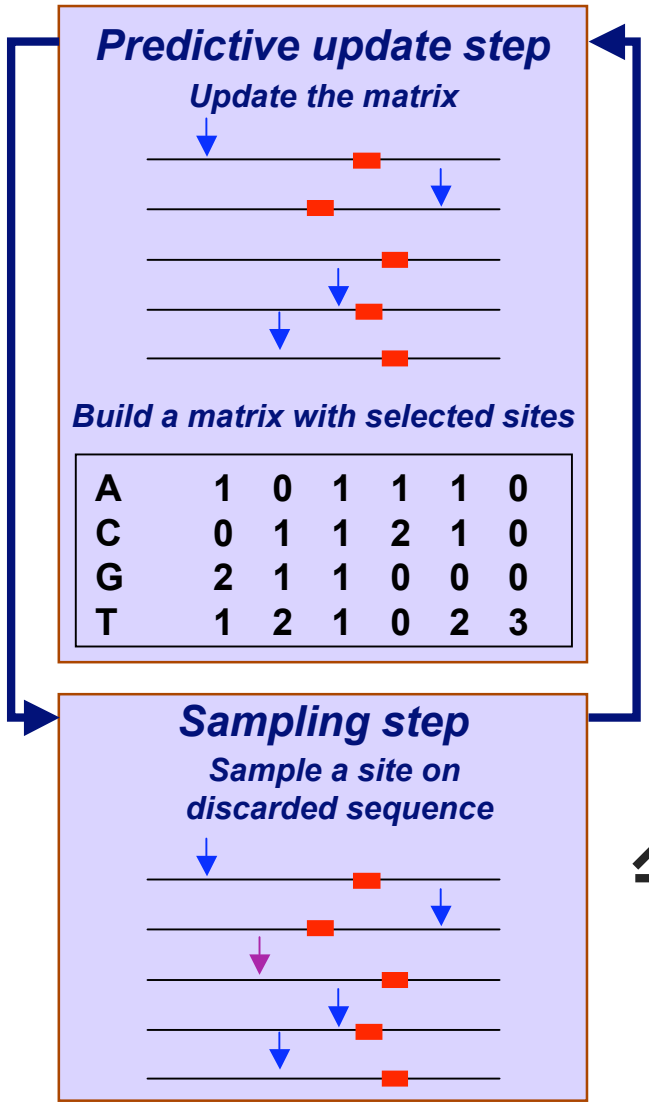
- Returns multiple matrices, but they are generally slight variants of the same pattern
- Time-consuming
- Sensitive to sequence ordering in the input data set
- Takes into account prior residue frequencies, but not oligonucleotide bias
- References
  - Hertz et al. (1990). Comput Appl Biosci 6(2), 81-92.
  - Hertz, G. Z. & Stormo, G. D. (1999). Bioinformatics 15(7-8), 563-77.
  - Stormo, G. D. & Hartzell, G. W. d. (1989). Proc Natl Acad Sci U S A 86(4), 1183-7.

Pattern discovery: The Gibbs sampler

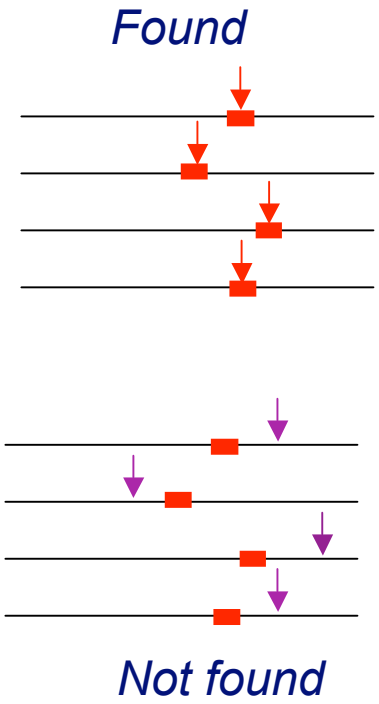
(gibbs motif sampler, by Andrew Neuwald)

Pretend you know the motif, this might become true

1) **Initialization**
   - o select a random set of sites in the sequence set
   - o Create a matrix with these sites
2) **Sampling**
   - o Isolate one sequence from the set, and score each site of the matrix
   - o Select one “random” site, with a probability proportional to the score (A_x, see next slide).
3) **Predictive update**
   - o Replace the old site with a new site, and update the matrix
4) **Iterate** steps 2 and 3 for a fixed number of cycles



After N iterations



Gibbs sampling - scoring scheme

$$A_x = Q_x / P_x$$

- A_x weight of segment x (used for random selection)
- Q_x probability to generate segment x according to pattern probabilities q_{ij}
- P_x probability to generate segment x according to the background probabilities p_i

$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + B}$$

- i index for the site
- j index for the residue
- $c_{i,j}$ counts for residue j at site i
- N number of sequences
- b_j pseudo-count for residue j
- B sum of pseudo-counts

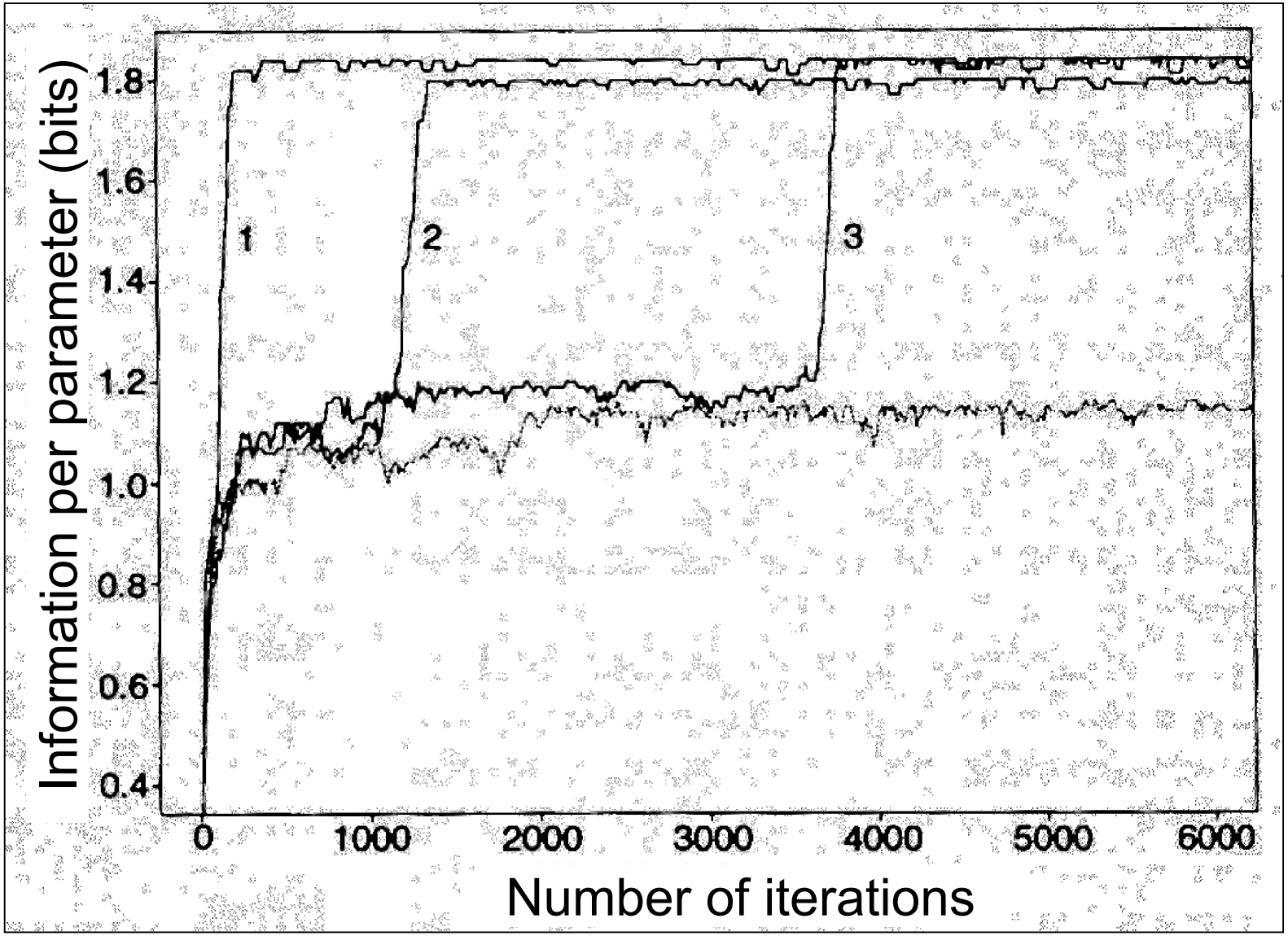
$$F = \sum_{i=1}^{W} \sum_{j=1}^{R} c_{i,j} \ln\left(\frac{q_{i,j}}{p_j}\right)$$

- W width of the matrix
- R number of distinct residues
- p_j prior probability for residue j

Stochastic vs deterministic behaviour

- Why to select a random site ?
  - A deterministic behaviour would consist in selecting, at each iteration, the highest scoring site (the one which matches best the matrix)
  - This would give poor results because the program is attracted too fast towards local optima.
- Stochastic behaviour
  - At each iteration, the next site is selected in a stochastic rather than deterministic way: the probability of each site to be selected is proportional to its scoring with the matrix
  - This allows to avoid weak local optima, and converge towards better solutions.

Gibbs sampling: optimization of information content



source: Lawrence et al.(1993). Science 262(5131), 208-14.

Gibbs sampling: strength

- Fast
- Probabilistic description of the patterns
- Can run with proteins or DNA

Gibbs sampling: weaknesses

- Returns a different result at each run
- Can be attracted by local maxima
  - solution: run repeatedly and check which motifs come often
- The original Gibbs sampler takes into account prior residue frequencies, but not oligonucleotide bias
  - → in yeast, often returns A/T-rich regions
  - This is however improved in some versions of the Gibbs samplers which use Markov chains for estimating the bacground probabilities (eg the MotifSampler developed by Gert Thijs)
- No threshold on pattern significance
  - → frequent false positive

Improvements of the gibbs sampler

- Neuwald 1993
  - Phase shifting
- Neuwald 1995
  - 0 or several matches per sequence
  - column sampling (spacings can be admitted between columns of the matrix)
- Roth (1998) : AlignACE
  - Specific implementation for DNA (double strand is treated)
  - post-filtering of motifs according to number of matches in the genome, in order to discard frequent motifs
- Lui (2000), Thijs (2000)
  - Markov-chain based calculation of background probabilities

References

- Original Gibbs sampler
  - Lawrence et al. (1993). Science 262(5131), 208-14.
  - Neuwald et al. (1995). Protein Sci 4(8), 1618-32.
  - Neuwald et al. (1997). Nucleic Acids Res 25(9), 1665-77.
- MotifSampler
  - Thijs et al. (2002). J.Computational Biology 9:447-464.

AlignACE, ScanACE and CompACE
gibbs sampler tools for regulatory sequence analysis

- Single/both strands
- Return multiple matrices, with iterative masking preventing slight variants of the same pattern
- Matrix clustering
- A posteriori evaluation of pattern significance, by analysing the whole-genome frequency of the discovered matrix.
- References
  - Roth et al. (1998). Nat Biotechnol 16(10), 939-45.
  - Tavazoie et al. (1999). Nat Genet 22(3), 281-5.
  - Hughes et al. (2000). J Mol Biol 296(5), 1205-14.
  - McGuire et al. (2000). Genome Res 10(6), 744-57.

Matrix-based pattern discovery: strengths

- More specific description of degeneracy than with string-based approaches (frequency of each residue at each position).
- The resulting pattern is more accurate than a string for pattern matching (more sensitive scoring scheme)

Matrix-based pattern discovery: weaknesses

- The results strongly depend on parameter setting. Two essential parameters have to be selected :
  - Matrix width
  - Expected number of sites
- The best parameter may change from gene family to gene family. Choosing the appropriate setting requires experience.
- Impossible to evaluate all possible alignments
- Does not take into account higher-order correlation between adjacent positions (oligonucleotide bias)