Regulatory Sequence Analysis

Matrix-based approaches for pattern discovery

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

Pattern discovery: typical dimensionality

- Typical case: GAL genes
  - s 6 sequences
  - L size per sequence 800 bp
  - occ_e expected pattern occurrences: 12 (2 sites per sequence)
  - w matrix width = 25
- Let us assume that
  - A signal can be found on any strand
  - Each sequence contains 0 or several occurrences
  - Number of possible site positions
    - $T=2s(L-w+1)=9312$

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

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

Regulatory Sequence Analysis

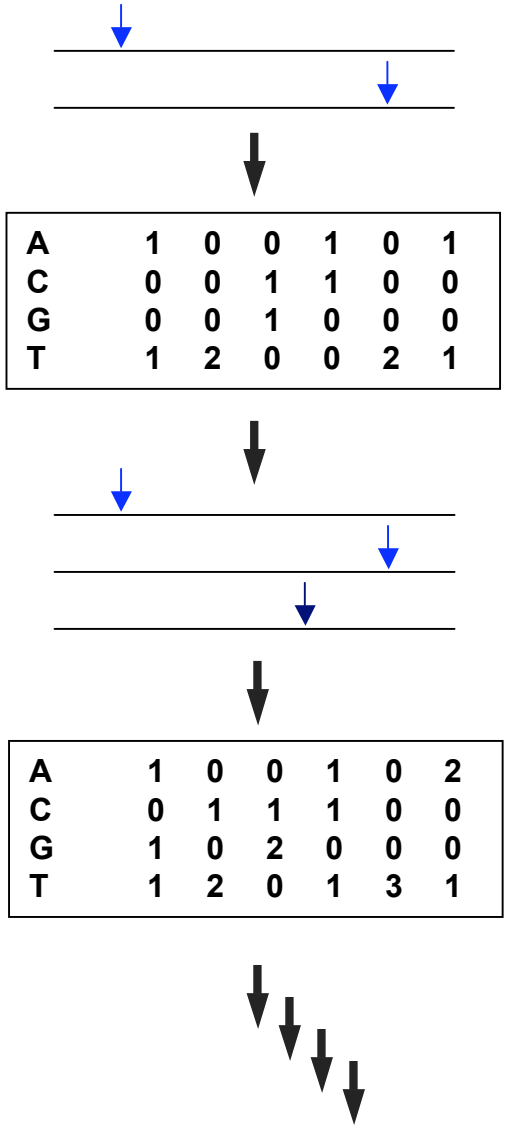
Greedy algorithm

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

Pattern discovery: greedy algorithm

(consensus, by Jerry Hertz)

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

Regulatory Sequence Analysis

Expectation- Maximization (EM)

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

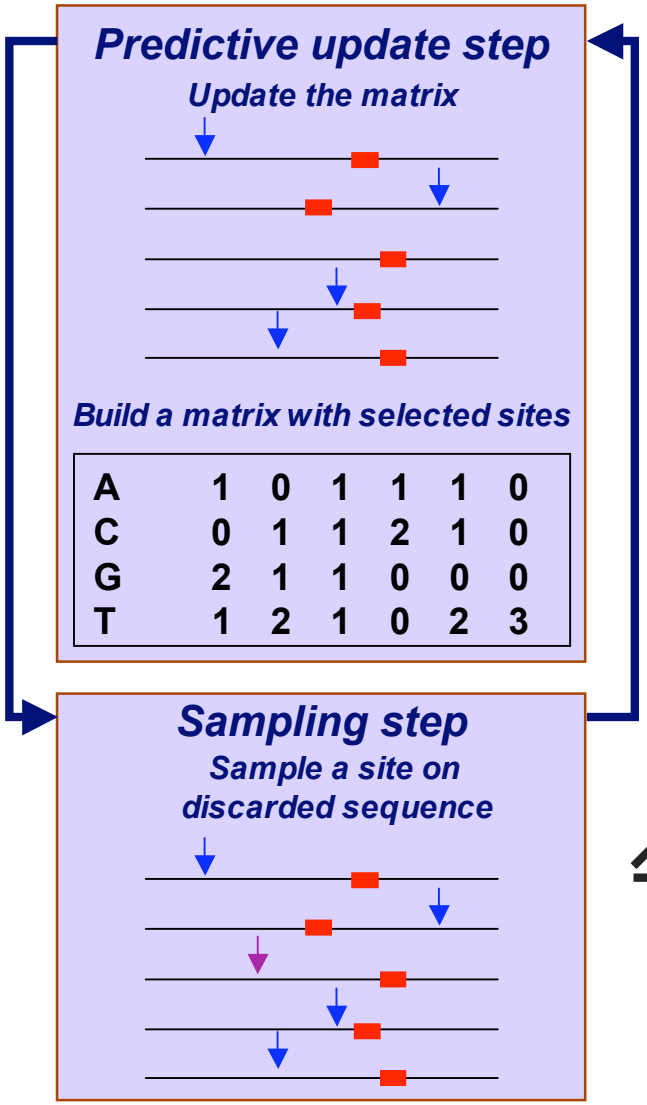
Gibbs sampling (stochastic Expectation - Maximization)

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

Pattern discovery: The Gibbs sampler (gibbs motif sampler, by Andrew Neuwald)

Pretend you know the motif, this might become true

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



After N iterations

Found

Not found

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

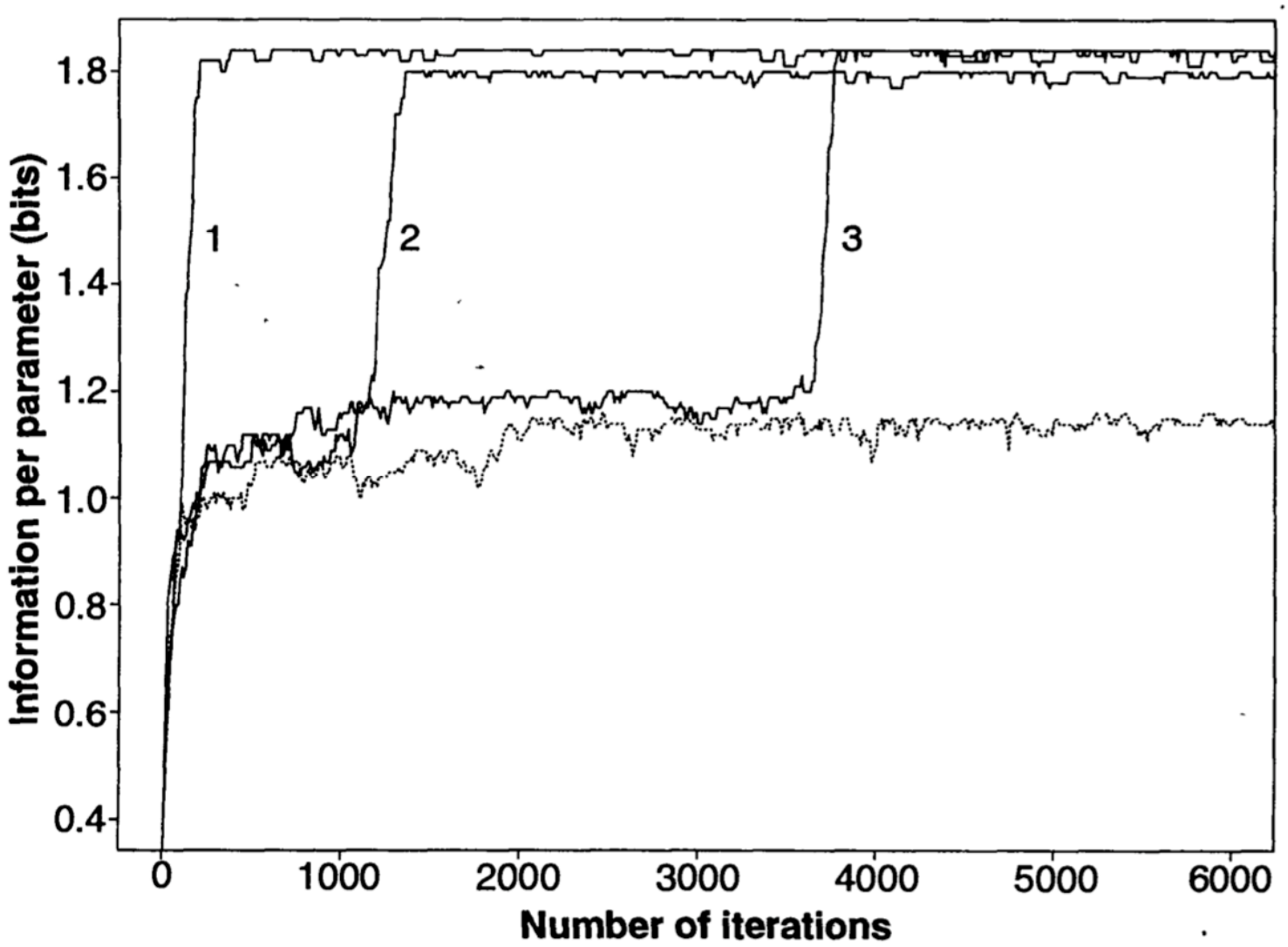
R number of distinct residues

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
- Liu (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)