Finding DNA sequence motifs and decoding *cis*-regulatory logic

Material from

- MacIsaac and Fraenkel, Practical strategies for discovering regulatory DNA sequence motifs, PLoS Comput Biol 2(4): e36
- Sacha et al Mechanisms and Evolution of Control Logic in Prokaryotic Transcriptional Regulation, Microbiol. Mol. Biol. Rev. 2009, 73(3): 481
- Das et al A primer on regression methods for decoding cis-regulatory logic, PLoS Comput Biol 5(1): e1000269

DNA sequence motifs

- Short recurring patterns in DNA that are presumed to have biological significance
- Often indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TFs)
- Others are involved in processes such as: ribosome binding, mRNA processing (splicing etc), transcription termination

How to find binding sites?

- Experimental: construct artificial sequences and explore binding affinities (using SELEX), Dnase footprinting
- Computational: search for overrepresented (and/or conserved) DNA patterns upstream of functionally related genes (e.g. genes with similar expression patterns or similar annotation)
- Huge gap between computational and experimental efforts
- Large-scale efforts to analyze genome-wide binding of TFs using ChIP-chip are rapidly addressing the gap
- Motif knowledge very useful in defining genetic regulatory networks and regulatory program of individual genes, so an important tool for computational biology

Regulation perspective: restriction enzymes

- Type II restriction enzymes that bind to DNA targets in highly specific sequence manner
- Part of a primitive bacterial immune system design to chop up viral DNA from infecting phages
- Cannot stray from consensus binding site => autoimmune reaction that could lead to irreversible damage to the bacterial genome
- Examples:
  - *Eco*RI binds to 6-mer GAATTC and only to that sequence
  - *Hin*dII binds to consensus sequence GTYRAC where Y stands for C or T (pYrimidine) and R stands for A or G (puRine)

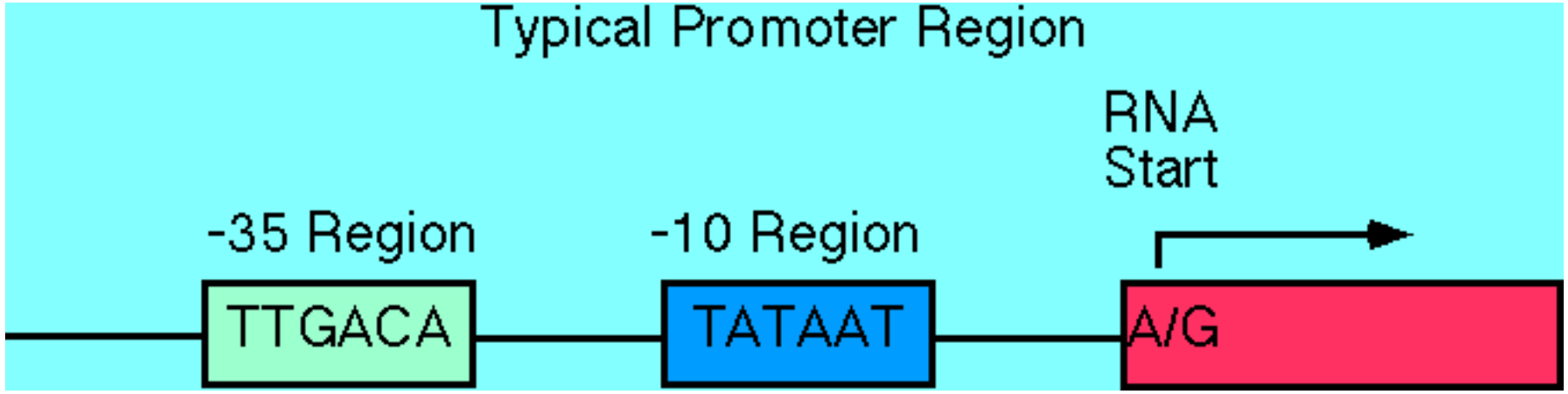
Consensus statistics

- Probability that a random 6-mer matches *Eco*RI binding site is (1/4)^6 so the site occurs about once every 4^6 = 4096 bp in a random DNA sequence

- For *Hin*dII however, there are two positions where two out of four bases can match, it would occur once per 4^4 x 2^2 = 1024 bp

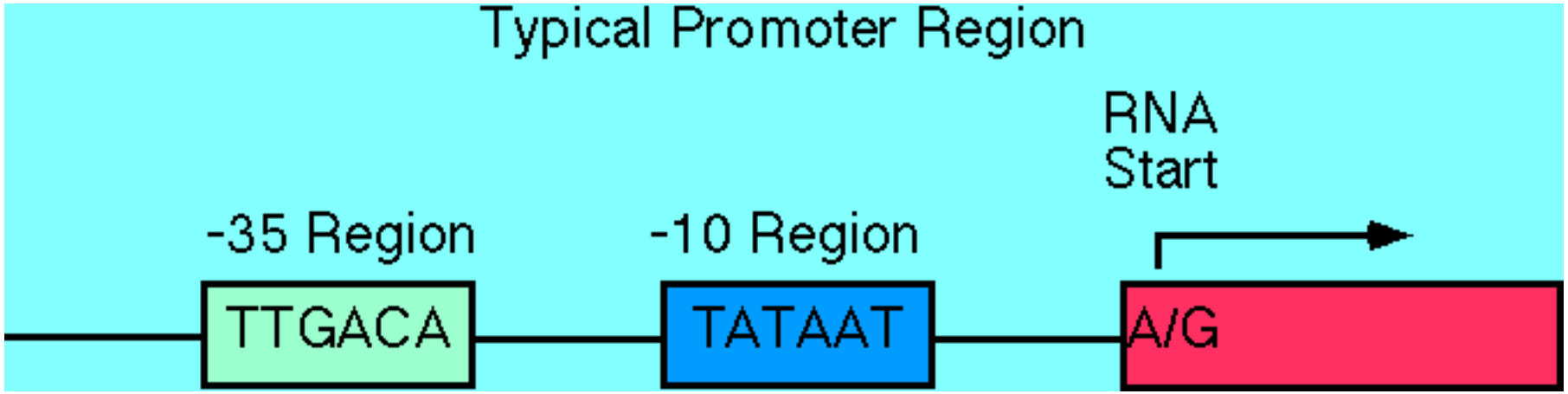
TATAAT box

- Well-conserved sequence centered around 10bp upstream of the transcription initiation site of *E coli* promoters
- Together with a motif TTGACA centered around -35, forms the binding for S70 subunit of the core RNA polymerase



TATAAT box

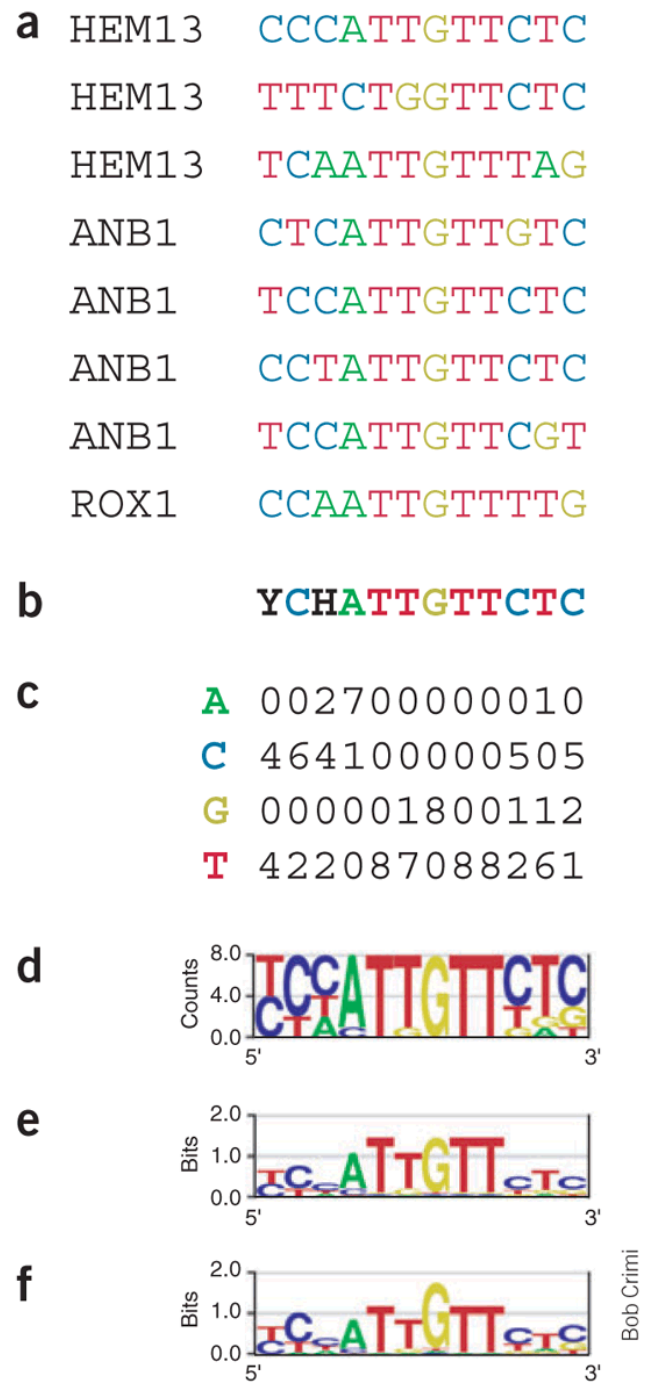
- Despite the high degree of conservation at each position (ranging from 54% to 82% for each base), it is rare to find a promoter that matches this consensus sequence exactly
- Most promoters match only 7-9 out of 12 bases



Position Weight Matrix (PWM)

- For TATAAT motif, activity of each promoter is related to how well it matches the consensus sequence, so the activity level of each gene can be fine-tuned by how much its -10 and -35 regions deviate from the consensus
- Use: **Position Weight Matrix (PWM)** to denote the fraction of nucleotide occurrences at each location of the motif and **Position Specific Scoring Matrix (PSSM)** to correct the occurrences for background distribution
- e.g. ROX1 transcription factor is known to bind at least 8 sites in three genes in the yeast (*Saccharomyces cerevisiae*) genome

- Panel a: multiple alignment of 8 binding sites of ROX1
- Consensus sequence in panel b. show a single base if it occurs more than half the sites and at least twice as often as the second most frequent base. Otherwise, use a double-degenerate symbol if two bases occur in more than 75% of the sties…
- Normalize columns in panel c to get PWM
- Core motif: ATTGTT



D'haeseleer, Nature Biotech 24, 4

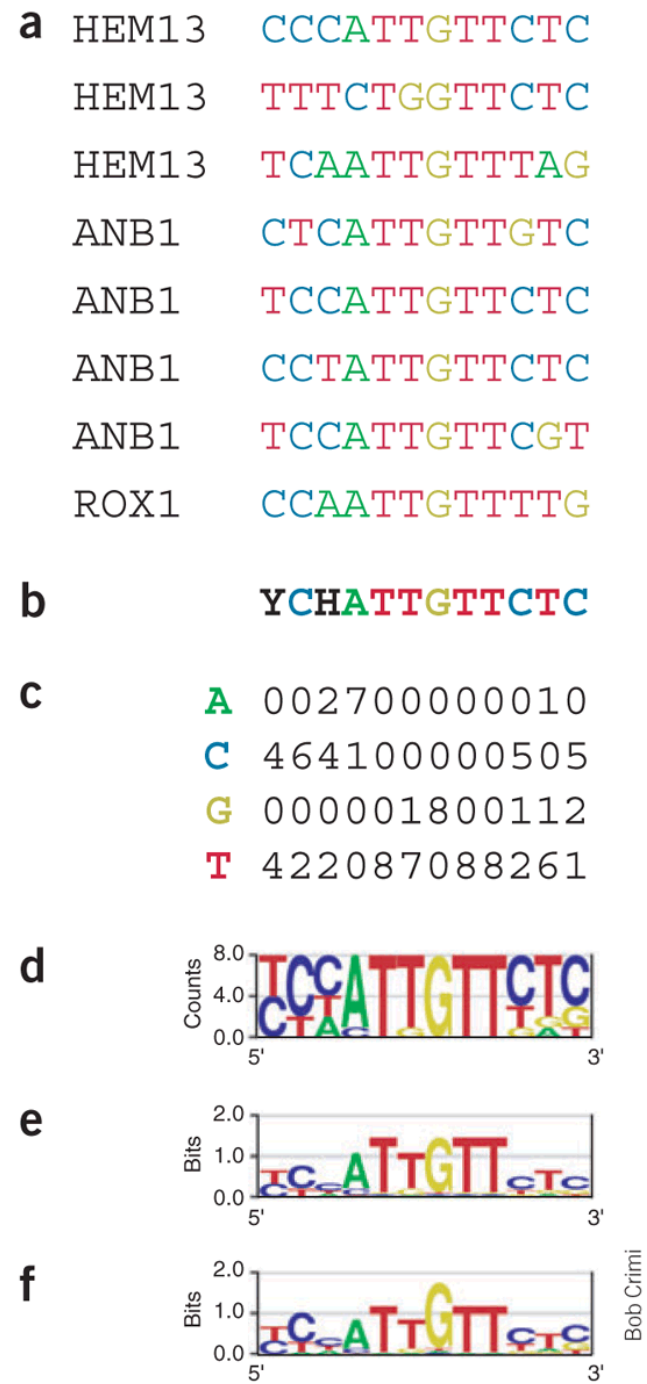
- Information content of a PWM :

$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

- $f_{b,i}$: frequency f of base b at position i

- Perfectly conserved: 2 bits
- Small sample corrections needed (panel e)

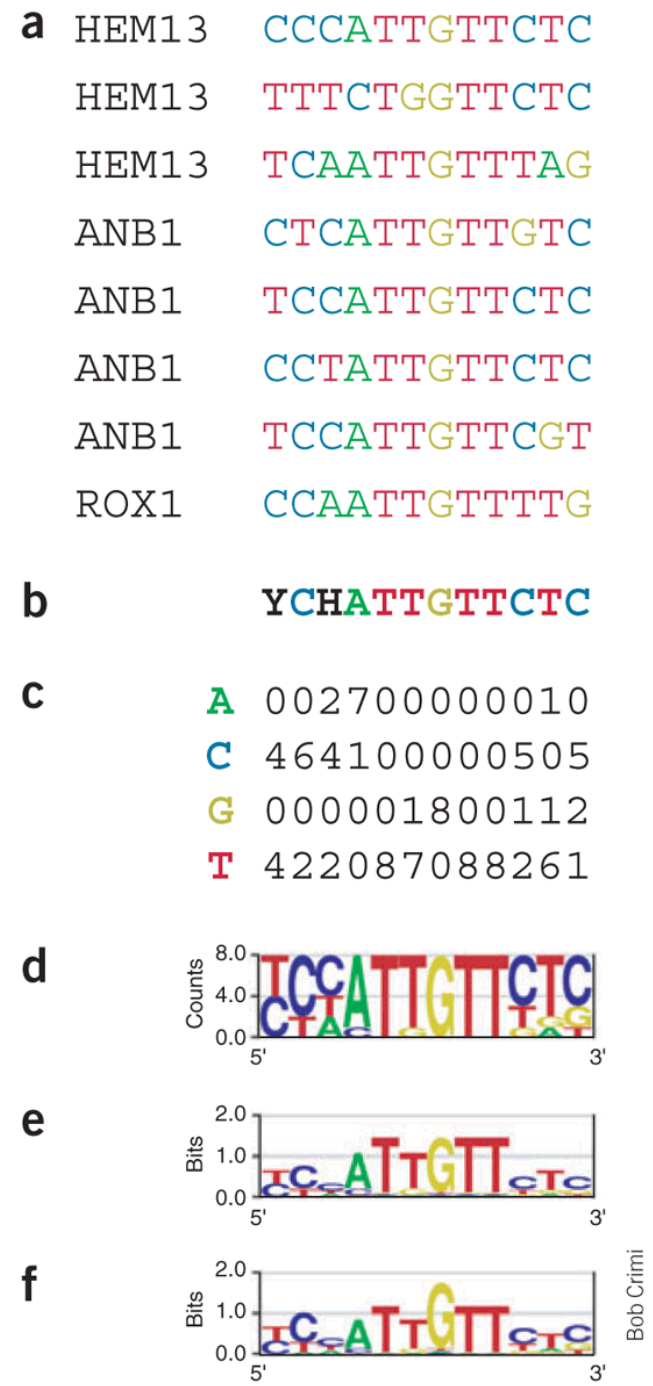
- Information content of partially degenerate 6-mer *Hin*dII: 10 bits



D'haeseleer, Nature Biotech 24, 4

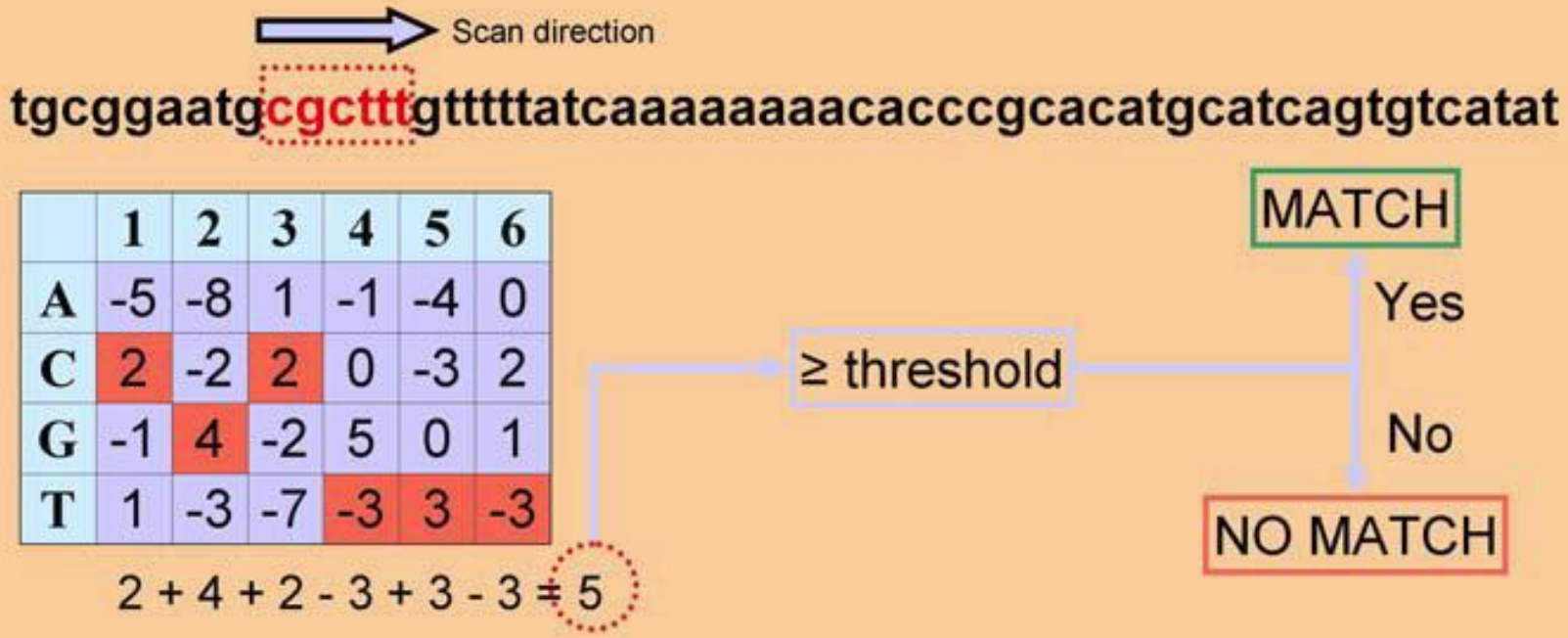
- **Position Specific Scoring Matrices (PSSM)**
- Correcting for background frequencies
  - All four bases occur equally is a reasonable approximation for *E. coli* (51% GC) or human (41% GC)
  - But is biased in *S. cerevisiae* (38%) *C. elegans* (36%), *Plasmodium falciparum* (19%), *Streptomyces coelicolor* (72%)
- Motif is interesting if it is different from the background distribution
- Use relative entropy (or information content) with base background frequency (panel f)

$$I_{seq}(i) = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$



D'haeseleer, Nature Biotech 24, 4

Position Weight Matrices define an additive scheme for scoring sequence. Often, the weights are simply log likelihood ratios of observing a nucleotide in a binding site relative to genomic background. Sequences are scanned by scoring every site, on both the forward and reverse complement strands, and identifying matches as shown in the schematic below:



A particular site is evaluated by adding up the entries from the scoring matrix at each position, and comparing the sum to a match threshold. For log ratio PWMs, an empirically chosen threshold of 60% of the maximum positive score has been used by Harbison et al. and is approximately equal to cutoffs determined by the principled cross-validated method presented in MacIsaac et al. More sophisticated algorithms developed specifically for motif scanning are described briefly in Figure 3.

TF information

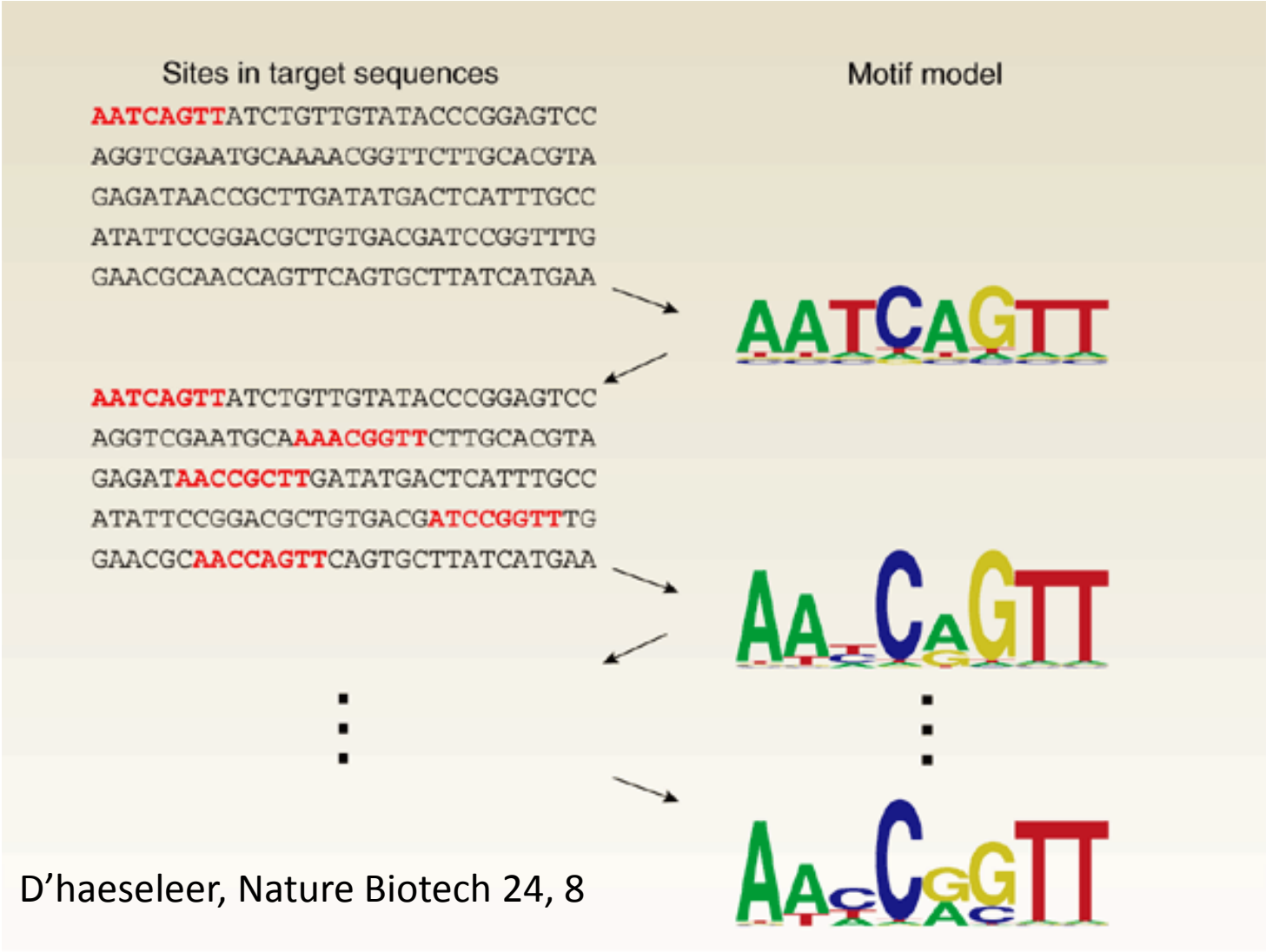
- TRANFAC
- JASPAR for multicellular eukaryotes
- YEASTRACT
- SCPD for *S. cerevisiae*
- RegulonDB for *E. coli*
- PRODORIC for prokaryotes

Motif discovery

- Three approaches:
  - Enumeration
  - Deterministic optimization
  - Probabilistic optimization
- Enumeration
  - Dictionary-based methods count # of occurrences of all n-mers in the target sequence and calculate which ones are overrepresented
  - Motif based description on exact occurrence is too rigid, use a flexible consensus description..or..
  - Search the space of all degenerate consensus sequences up to a given length
  - Use a consensus sequence and allow mismatches, use suffix tree representation to find all such motifs in target sequences
  - No getting stuck in local minima, but these methods may overlook some of the subtle patterns present in the real binding sites

Deterministic optimization

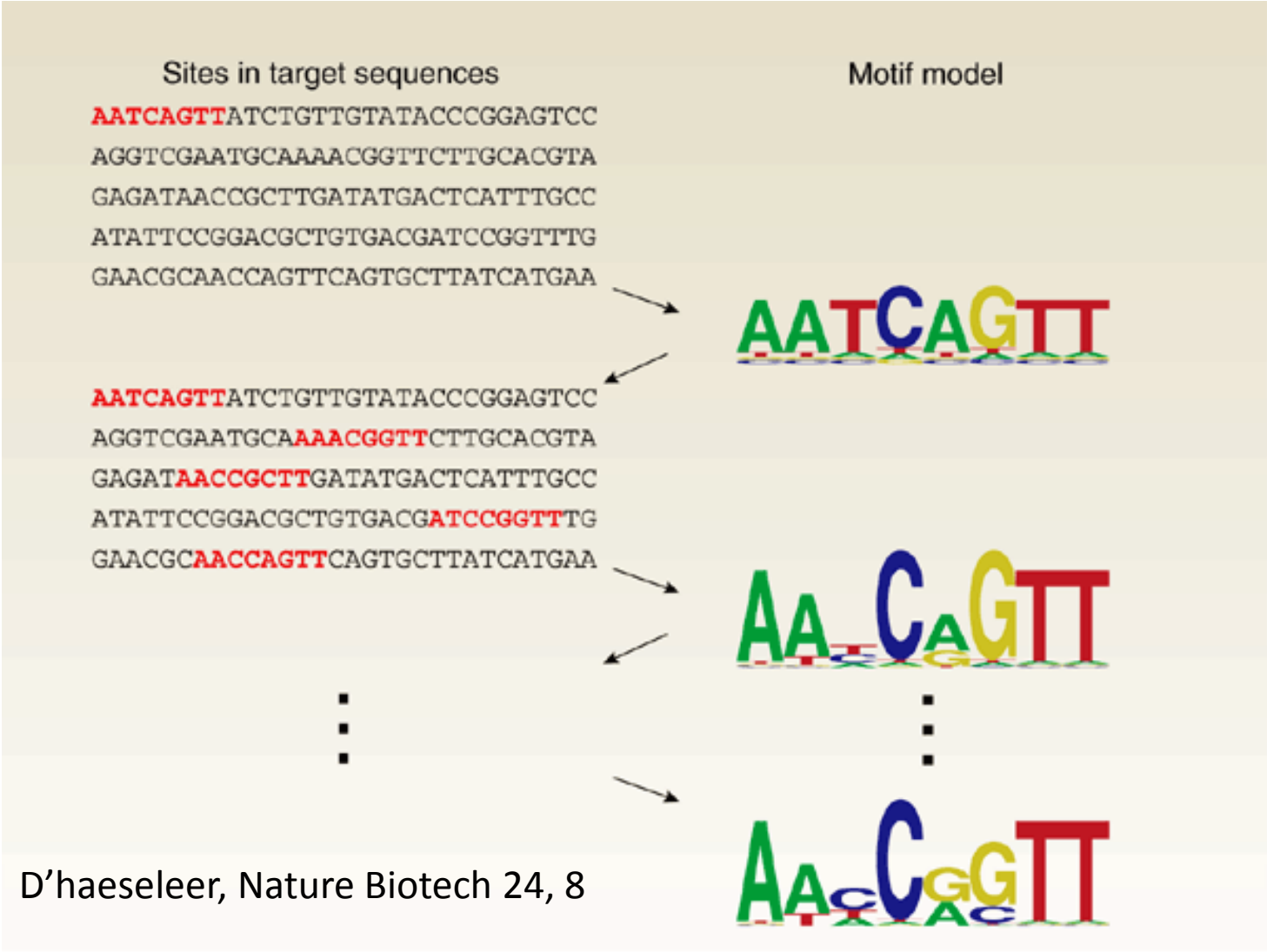
- Use EM to simultaneously optimize a position weight matrix (PWM) description of a motif and the binding probabilities for tis associated sites
- Initialize the weight matrix for the motif with a single n-mer subsequence plus a small amount of background nucleotide target sequences
- For each n-mer in the target sequence, calculate the probability that it was generated by the motif, and compare that with the probability assigned by a background sequence distribution
- EM takes a weighted average across these probabilities to generate a more refined motif model
- Algorithm iterates between calculating the probability of each site based on the current motif model and calculating a new motif based on the probabilities



D'haeseleer, Nature Biotech 24, 8

Deterministic optimization

- MEME: multiple EM for motif elicitation
- MEME performs a single iteration for each n-mer in the target sequences, selects the best motif from this set and then iterates only that one to convergence, avoiding local maxima
- Find additional motifs by masking the sequences matched by the first motif and rerunning the algorithm



D’haeseleer, Nature Biotech 24, 8

Probabilistic Optimization

- Gibbs sampling: stochastic implementation of EM
- Initialize motif model with randomly selected set of sites
- Every site in the target sequence is scored against this initial motif model
- At each iteration, probabilistically decide whether to add a new site and/or remove an old site from the motif model, weighted by the binding probability of these sites
- Update the resulting motif model and recalculate the binding probabilities
- After many iterations, we would have sampled the joint probability distribution of motif models and sites assigned to the motif, focusing in on the best fitting combinations

Which one to use?

- Tompa et al compared 13 different motif discovery algorithms
  - Enumerative approaches: Weeder and YMF performed well on eukaryotic sequences with known motifs
  - Each algo covered only a small subset of known binding sites, with relatively little overlap between the algorithms
  - Best to combine results from multiple discovery tools: MotifSampler
  - Implementation details may be more important than optimization procedure
    - How to represent motifs
    - Whether to optimize motif width and number of occurrences
    - Which objective function

Binding energy and searching for new sites

- Affinity of a DNA binding protein to a specific binding site is typically correlated with how well the site matches the consensus sequnces
- But not all matches in a binding site are equally forgiving of mismatches and not all matches at a given position have the same effect
- Assume each position contributes to the binding energy independently, we could measure the effect of binding energy of all possible base changes
- The resulting PWM, call it $W(b,i)$, can be used to calculate the specific-binding free energy (relative to random background DNA) of a sequence S, where $S(i)$ is the base occurring in position i in sequence S:

$$-\Delta G_s(S) = \sum_i W(S(i),i)$$

Biophysical interpretation

- We usually have a list of known binding sites, without any affinity information.
- If we assume that the genomic DNA is random with base frequencies pb, we can optimize the values of PWM such that the probability of binding to the known binding sites (versus the more abundant background DNA) is maximized.
- Optimal weight matrix:

$$W(b,i) = \log_2 \frac{f_{b,i}}{p_b}$$

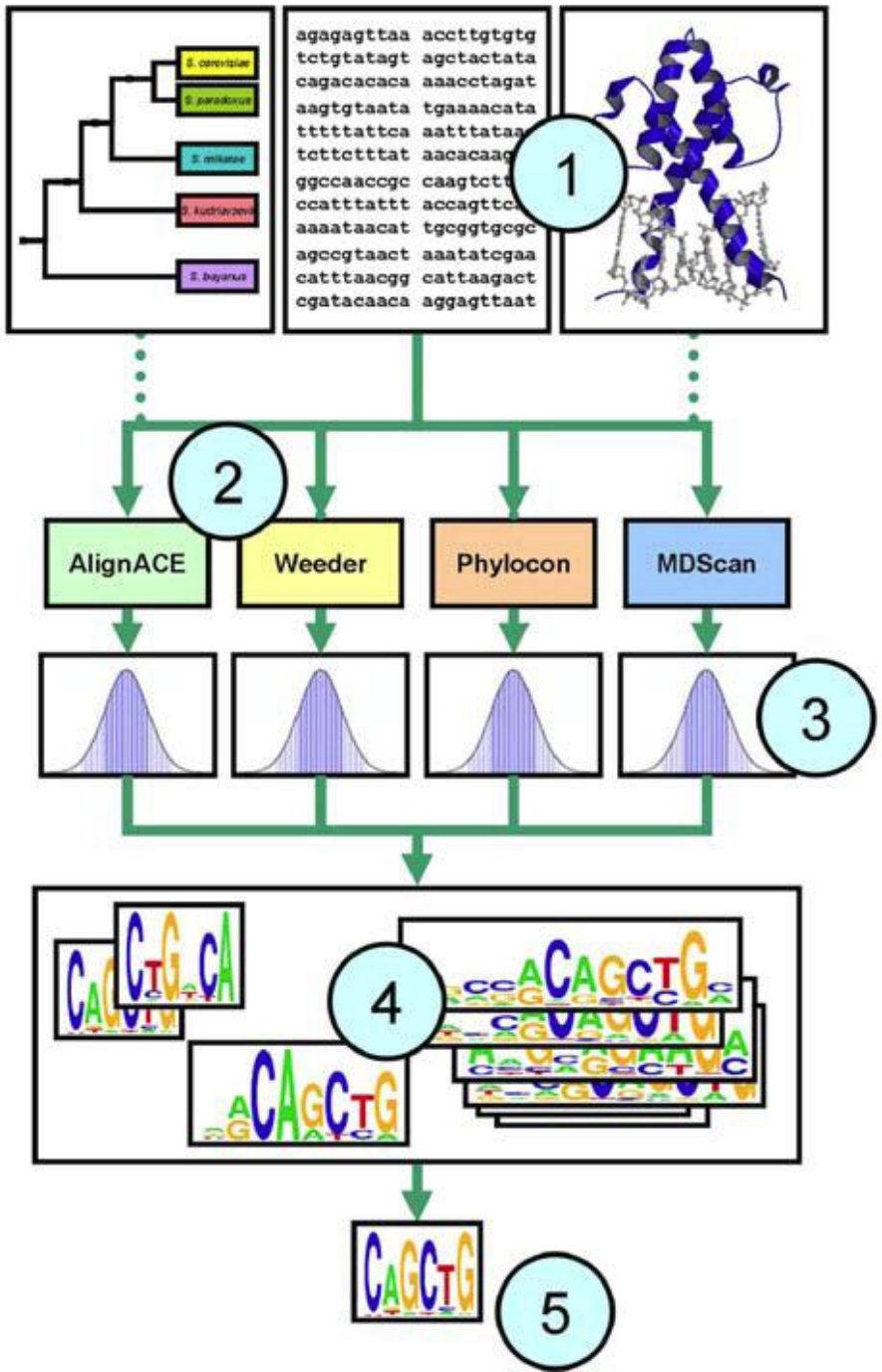
- Information content of a sequence can be interpreted as an estimate of the average specific binding energy to the entire set of known binding sites, in competition with genomic DNA

Which one of the motifs is biologically relevant?

- Information content
- Lo-likelihood
- MAP score
- Group specificity: probability of having this many target sequences containing the site (or this many sites within the target sequences), considering the prevalence of the motif throughout the genome
- Sequence specifity: emphasize both the number of sequences with binding sites, and the number of sites per sequence
- Positional bias or uniformity: real TF binding sites often (but not always) show a marked preference for a specific region upstream of the genes they regulate. So measure how uniform the binding site locations are distributed with respect to transcription start site of the gene.
- Experimental: phylogenetic footprinting and ChIP-chip analysis

Guidelines

- If possible, remove spurious patterns from target sequences (RepeatMasker)
- Use multiple motif prediction algorithms
- Run probabilistic algorithms many time
- Retrieve multiple motifs
- Try a range of motif widths and expected number of sites
- Filter out motifs with biologically implausible distribution of information content (‘block filtering’)
- Combine similar motifs, AlignACE, cluster and take best representative
- Use AlignACE to match up with known motifs for the organism
- Evaluate resulting motifs with criterion on the previous page



1. **Assemble input data.** Results may be improved by restricting the input to high-confidence sequences. Some algorithms achieve improved performance by using phylogenetic conservation information from orthologous sequences or information about protein DNA-binding domains.

2. **Choose several motif discovery programs for the analysis.** For recommended programs see Figure 3.

3. **Test the statistical significance of the resulting motifs.** Use control calculations to estimate the empirical distribution of scores produced by each program on random data.

4. **Clustering and post-processing the motifs.** Motif discovery analyses often produce many similar motifs, which may be combined using clustering. Phylogenetic conservation information may be used to filter out statistically significant, but non-conserved motifs that are more likely to correspond to spurious sequence patterns.

5. **Interpretation of motifs.** Algorithms exist for linking motifs to transcription factors and for combining motif discovery with expression data.

Multi-purpose packages

TAMO	TAMO integrates several motif discovery programs. It includes support for motif scanning, scoring, evaluation of statistical significance, clustering, comparison, input/output, conversion between different motif representations, and visualization. http://fraenkel.mit.edu/webtamo/
BEST	BEST is a suite of four motif discovery tools integrated in a graphical user interface. BEST incorporates the BioOptimizer tool used to rank and improve the predictive power of the discovered motifs. http://webster.cs.uga.edu/~che/BEST/
TOUCAN2	TOUCAN2 provides an interface to the Ensembl and EMBL databases of sequence and annotation. It incorporates tools for sequence alignment, motif discovery, and scanning. http://homes.esat.kuleuven.be/~saerts/software/toucan.php
Expander	Expander is a tool for analyzing expression data. It can cluster genes, identify over-represented functional categories in clusters, and scan corresponding promoter regions for motifs. http://www.cs.tau.ac.il/~rshamir/expander/
MDScan **BioProspector** **Compare-Prospector**	MDScan uses ChIP-chip enrichment ratio data to help the motif search. BioProspector is a Gibb's sampling program. CompareProspector incorporates comparative genomics, biasing the search to regions of high conservation. http://seqmotifs.stanford.edu
Consensus **PhyloCon**	The Consensus program finds motifs in a set of unaligned sequences. PhyloCon builds on this framework by modeling conservation across orthologous genes from multiple species. http://ural.wustl.edu/
Weeder	An enumerative motif discovery program that performed well in a recent comparative analysis of fourteen algorithms. http://www.pesolelab.it/
MEME	The popular EM-based motif discovery program. Part of the MEME/MAST system for motif discovery and search. http://meme.sdsc.edu/meme/intro.html
AlignACE	A Gibbs sampling algorithm that can identify multiple motifs in a sequence set using an iterative masking procedure. http://atlas.med.harvard.edu/

Motif Scanning

Ahab	The Ahab webserver allows users to scan for motifs in a set of sequences. Motifs may be user-specified or selected from a database of pre-defined matrices. http://gaspard.bio.nyu.edu/Ahab.html
Clover	Clover identifies overrepresented motifs in a set of sequences, based on a pre-compiled library of motif matrices. http://zlab.bu.edu/clover/
MAST	MAST allows users to scan sequence databases for matches to motifs. It produces detailed annotations and figures for matches in the input sequences. http://meme.sdsc.edu/meme/intro.html
Monkey	Monkey analyzes multiple sequence alignments to identify evolutionarily conserved matches to a motif. http://rana.lbl.gov/~alan/Monkey.htm

Motif Discovery Programs

Databases

cisRED **ORegAnno**	cisRED is a database of conserved motifs and motif patterns obtained by genome scale motif discovery. ORegAnno is a database of regulatory sites curated from the scientific literature. http://www.cisred.org/ http://www.oreganno.org/
UCSC Genome Browser	Online repository of genomic sequence, multiple sequence alignments, and annotation data. The browser includes tracks for identifying conserved transcription factor binding sites. http://genome.ucsc.edu/
ENSEMBL	Another online genomic sequence repository. Includes online tools for data mining as well as BLAST searches. http://www.ensembl.org/index.html
TRANSFAC	Commercial database of transcription factors, binding sites, and motifs. Includes several tools for motif scanning in sequence. http://www.gene-regulation.com/
JASPAR	Curated public database of transcription factor binding specificities represented as PWMs. http://jaspar.cgb.ki.se/

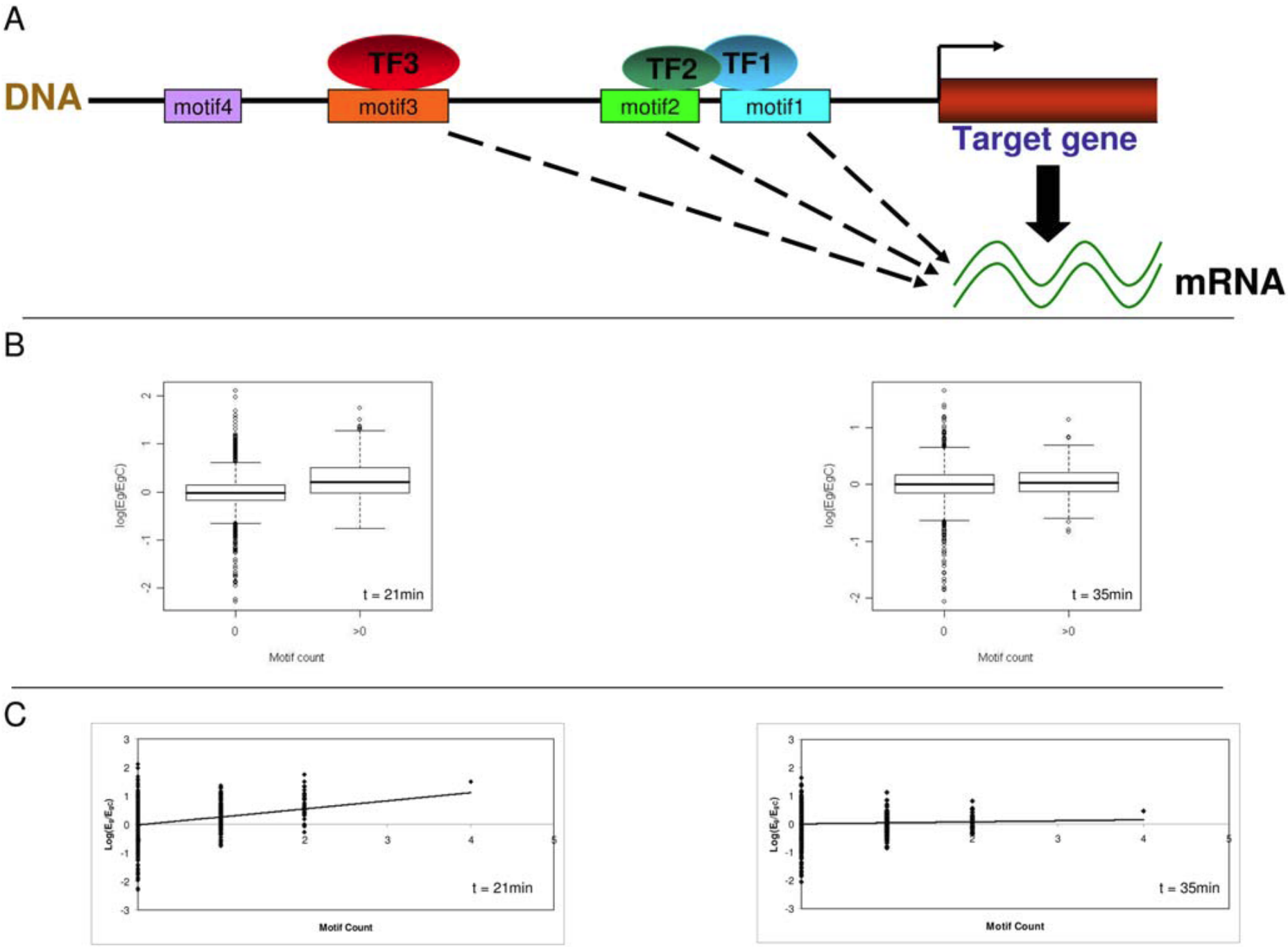
Searching for new motifs with biophysics

- The binding energy PWM can be used to search for novel sites, using a scoring threshold based on scores of known binding sites
- False positives?
- Do simultaneous optimization of weight matrix and thresholds
- But for large eukaryotic genomes, expect low affinity hits.
- Other factors: chromatin structure, cooperative binding must play a role in determining in vivo specificity of associated TFs

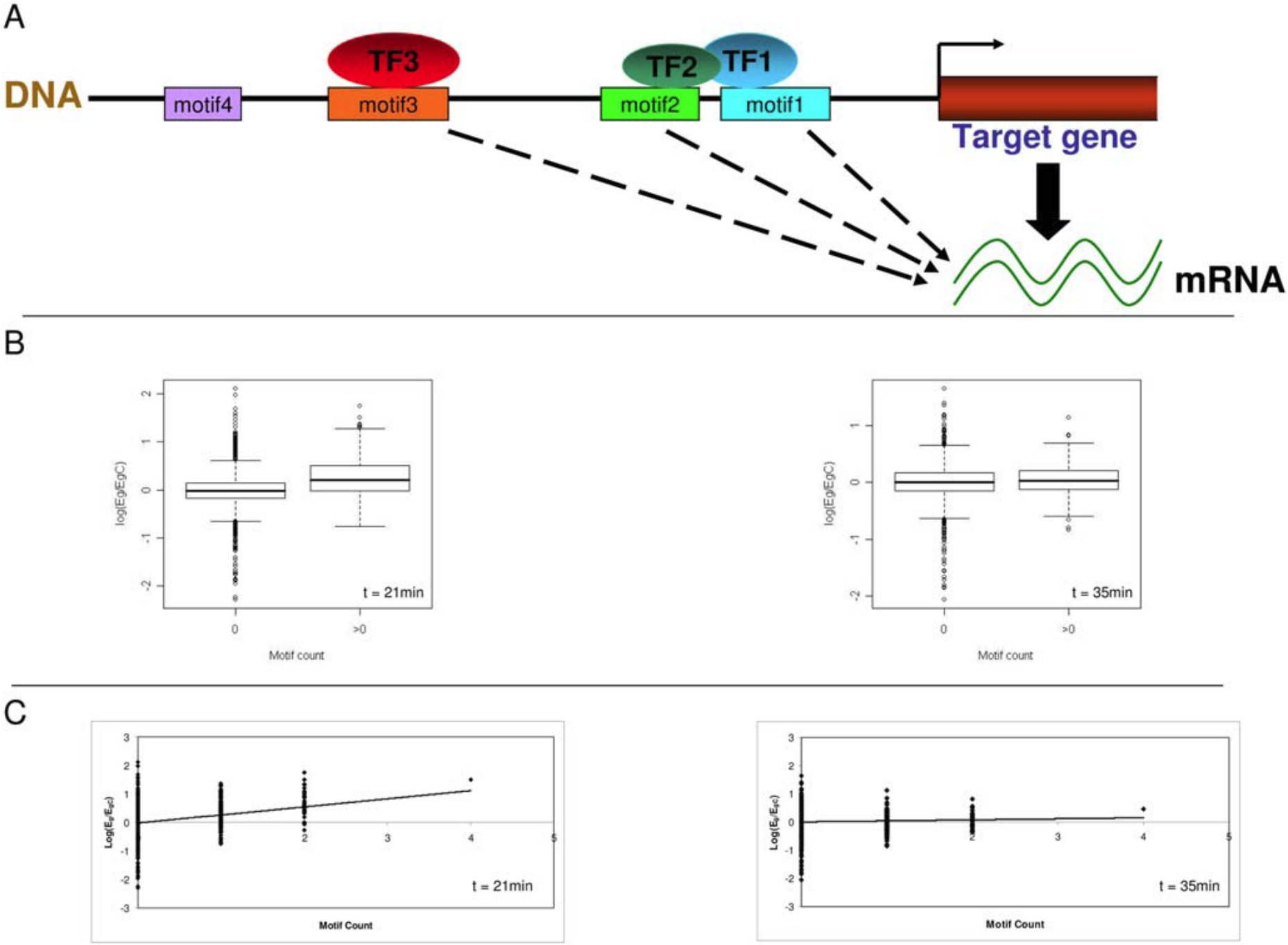
Cis regulatory module (CRM)

- Stretch of DNA, usually 100-1000 DNA bp in length
- # of TFs can bind and regular expression of nearby genes
- Cis because they are typically located on the same DNA as the genes they control, as oppoed to trans, which refers to effects on genes not located on the same strand or farther away
- One cis-regulatory element can regular several genes
- One gene can have several cis-regulatory modules

- Motifs 1, 2 and 3 are bound to TFs and thus are active
- Motif 4 is not
- TFs 1 and 2 are shown to be interacting
- Experiment on yeast-cell cycle
- Look for gene expression to know which elements are active



- MCB element ACGCGT regulates G1/S phase
- Expression changes for G1/S but not G2/M
- Box plots: log (Eg/Egc) for genes in two different phases
- Linear regression models



- Non-linear regression models