Revisar principalmente la parte resaltada

What are DNA sequence motifs?

Patrik D'haeseleer

Sequence motifs are becoming increasingly important in the analysis of gene regulation. How do we define sequence motifs, and why should we use sequence logos instead of consensus sequences to represent them? Do they have any relation with binding affinity? How do we search for new instances of a motif in this sea of DNA?

Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function. Often they indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TF). Others are involved in important processes at the RNA level, including ribosome binding, mRNA processing (splicing, editing, polyadenylation) and transcription termination.

In the past, binding sites were typically determined through DNase footprinting, and gel-shift or reporter construct assays, whereas binding affinities to artificial sequences were explored using SELEX. Nowadays, computational methods are generating a flood of putative regulatory sequence motifs by searching for overrepresented (and/or conserved) DNA patterns upstream of functionally related genes (for example, genes with similar expression patterns or similar functional annotation). For a while, it seemed like we had more computationally predicted sequence motifs without a known matching transcription factor, than transcription factors without a known binding sequence, although large-scale efforts to analyze the genome-wide binding of transcription factors using ChIP-chip are rapidly rectifying this situation.

The abundance of both computationally and experimentally derived sequence motifs and their growing usefulness in defining genetic regulatory networks and deciphering the regulatory program of individual genes make them important tools for computational biology in the post-genomic era.

Patrik D'haeseleer is in the Microbial Systems Division, Biosciences Directorate, Lawrence Livermore National Laboratory, PO Box 808, L-448, Livermore, California 94551, USA
e-mail: patrikd@llnl.gov

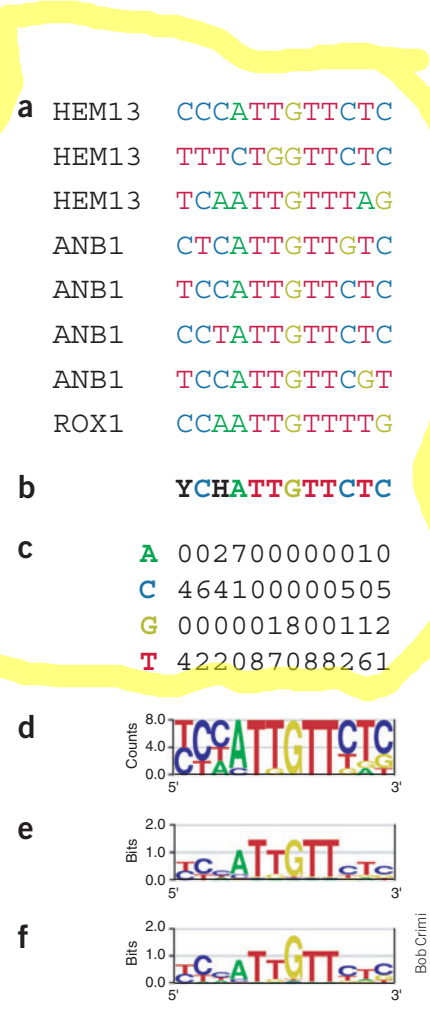

Figure 1 ROX1 binding sites and sequence motif. (**a**) Eight known genomic binding sites in three *S. cerevisiae* genes. (**b**) Degenerate consensus sequence. (**c**,**d**) Frequencies of nucleotides at each position. (**e**) Sequence logo showing the frequencies scaled relative to the information content (measure of conservation) at each position. (**f**) Energy normalized logo using relative entropy to adjust for low GC content in *S. cerevisiae*.

Restriction enzymes and consensus sequences

Type II restriction enzymes, discovered in the late 1960s, need to bind to their DNA targets in a highly sequence-specific manner, because they are part of a primitive bacterial immune system designed to chop up viral DNA from infecting phages. Straying from their consensus binding site specificity would be the equivalent of an autoimmune reaction that could lead to irreversible damage to the bacterial genome. For example, *Eco*RI binds to the 6-mer GAATTC, and only to that sequence. Note that this motif is a palindrome, reflecting the fact that the *Eco*RI protein binds to the DNA as a homodimer. Other restriction enzymes bind to a degenerate consensus sequence. For example, *Hin*dII bind to the sequences GTYRAC, where Y stands for 'C or T' (pYrimidine), and R stands for 'A or G' (puRine). (See http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html#tab1 for a listing of the IUPAC symbols for degenerate consensus sequences.)

We can calculate how often we would expect these consensus sequences to occur, based on their length and degeneracy. The probability that a random 6-mer matches the *Eco*RI binding site is $(1/4)^6$, so the site occurs about once every 4^6 (= 4,096) bp in a random DNA sequence. The *Hin*dII binding site, containing two positions where two out of four bases can match, would occur once per $4^4 \times 2^2$ (= 1,024) bp.

Consensus or caricature?

Other DNA binding proteins tend to be less picky in sequence specificity. In 1975, Pribnow discovered the 'TATAAT box,' a well-conserved sequence centered around 10 bp upstream of the transcription initiation site of *Escherichia coli* promoters. This motif, together with a

TTGACA motif centered around –35, forms the binding site for the σ^{70} subunit of the core RNA polymerase. However, despite the high degree of conservation at each position (ranging from 54% to 82% for each base), it is actually extremely rare to find a promoter that matches this consensus sequence exactly, with most promoters matching only 7–9 out of the 12 bases. Rather than representing a typical binding sequence, the consensus sequence in this case is instead a highly unusual sequence. It turns out that the 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.

A better description of the binding sequence in this case is through a Position Frequency Matrix (PFM). Rather than only keeping track of the most common base at each position, we record how often each base occurs in known sites. For example, the Rox1 transcription factor is known to bind at least eight sites in three genes in the *Saccharomyces cerevisiae* genome. **Figure 1** shows the multiple alignment of these eight binding sites, with a consensus sequence of YCHATTGTTCTC. (Conventionally, a single base is shown if it occurs in more than half the sites and at least twice as often as the second most frequent base. Otherwise, a double-degenerate symbol is used if two bases occur in more than 75% of the sites, or a triple-degenerate symbol when one base does not occur at all.) The frequency matrix and its graphic representation in **Figure 1** clearly show a core motif of ATTGTT, with much lower conservation in the flanking bases.

Sequence logos

By scaling each stack of letters in **Figure 1d** with some measure of the conservation at each base, we get a much clearer view of the binding sequence. In a 'sequence logo,' developed by Schneider and Stephens[1], each stack is scaled with the information content of the base frequencies at that position:

$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i} \qquad (1)$$

where $f_{b,i}$ indicates the frequency of base b at position i. Positions that are perfectly conserved contain 2 bits of information, those where two of the four bases occur 50% of the time each contain 1 bit, and positions where all four bases occur equally often contain no information. Note that for a small sample, the information content will tend to be overestimated, so a small-sample correction needs to be applied. This explains why the central positions of the motif in **Figure 1e** show an information content of less than 2 bits, even though they are perfectly conserved within the eight known binding sites.

Note that the total information content of a motif is directly related to its expected frequency of occurring within a random DNA sequence. For example, the information content of the partially degenerate 6-mer *Hin*dII binding site is 10 bits (2 bits per conserved base, 1 bit per double-degenerate position), and its expected frequency in random DNA is 1 in $2^{10} = 1{,}024$.

Correcting for background frequencies

Equation (1) assumes all four bases occur equally often in the background genomic DNA. For organisms such as *E.coli* (51% GC) or human (41%) this is usually a reasonable approximation. However, for genomes with a more biased GC content such as *S. cerevisiae* (38%), *Caenorhabditis elegans* (36%) and especially extremes such as *Plasmodium falciparum* (19%) or *Streptomyces coelicolor* (72%), a correction factor is needed. One approach—advocated by Schneider—is to replace the '2' in equation (1) with the lower entropy of random DNA of the specified GC content. A more informative approach[2] is to generalize equation (1) to the relative entropy (a.k.a. Kullback-Leibler distance) of the binding site with respect to the background frequencies:

$$I_{seq}(i) = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b} \qquad (2)$$

where p_b is the background frequency of base b in the genome. This is equivalent to a log-likelihood ratio (G test) to measure the degree of disagreement between the observed and background base frequencies, and thus can again be used to calculate the significance of the motif itself (and the frequency of occurrence of such a sequence in random DNA).

Figure 1f shows the Rox1 binding motif, corrected for the GC-content of *S. cerevisiae* genomic DNA using equation (2). In comparison with **e**, the central G base now carries more information than the flanking A and T bases, reflecting the fact that its occurrence is much more significant in the low-GC genome. The total information content of the motif is $I_{seq} = 11.27$ bits.

Roll your own logos

Two free web interfaces exist for generating a sequence logo from your favorite DNA alignment: Steven Brenner's WebLogo (http://weblogo.berkeley.edu/), implementing Schneider's original Sequence Logos, and the more recent enoLOGOS[3] (http://biodev.hgen.pitt.edu/enologos), using relative entropy. The former provides an option to put error bars on the information content, which can be quite useful especially for motifs based on a small number of sequences. However, the latter offers a wider variety of input formats, variable GC content, and the option to examine nonindependent bases via mutual information. The two sites also take a different approach to small-sample correction. The logos in **Figure 1** were generated using enoLOGOS.

Transcription factor binding sites are collected in a number of online databases, including TRANSFAC[4] (http://www.generegulation.com/pub/databases.html), JASPAR[5] for multicellular eukaryotes (http://jaspar.genereg.net), YEASTRACT[6] (http://www.yeastract.com/) and SCPD[7] (http://rulai.cshl.edu/SCPD) for *S. cerevisiae*, RegulonDB[8] for *E. coli* (http://regulondb.ccg.unam.mx) and PRODORIC[9] for prokaryotes (http://www.prodoric.de/), although some of these are still focused primarily on consensus sequences.

Binding energy and searching for novel sites

As mentioned above, the affinity of a DNA binding protein to a specific binding site is typically correlated with how well the site matches the consensus sequence. However, not all positions in a binding site are equally forgiving of mismatches, and not all mismatches at a given position have the same effect.

If we assume that each position contributes to the binding energy independently (a reasonable approximation in most cases), we could laboriously measure the effect on binding energy of all possible single base changes. The resulting Position Weight Matrix (PWM) $W(b,i)$ can then be used to calculate the specific-binding free energy (relative to random background DNA) of a sequence S as:
where $S(i)$ is the base occurring in position i in sequence S.

$$-\Delta G_s(S) = \sum_i W(S(i),i) \qquad (3)$$

Typically, we only have a list of known binding sites, without any affinity information. If we assume that the genomic DNA is random with base frequencies p_b, it is possible to optimize the values in the PWM such

that the probability of binding to the known binding sites (versus the more abundant background DNA) is maximized. The optimal weight matrix is then given by:

$$W(b,i) = \log_2 \frac{f_{b,i}}{p_b} \quad (4)$$

The information content I_{seq} can then be interpreted as an estimate of the average specific binding energy to the entire set of known binding sites, in competition with the genomic DNA.

This PWM can be used to search for novel sites with high predicted binding affinity within the rest of the genome, typically using a score threshold based on the scores of the known binding sites. Unfortunately, this approach can result in large numbers of false positives, sometimes returning hundreds or thousands of putative binding sites. A promising alternative developed by Djordjevic *et al.*[10] is to simultaneously optimize the weight matrix as well as the threshold such that all the known sites are included, but as few other sites as possible, typically resulting in many fewer and more reliable novel sites.

Nevertheless, it is unavoidable that sequence motifs with low information content (that is, short motifs, and/or with a lot of degeneracy) will tend to yield large numbers of fairly low affinity hits, especially in large eukaryotic genomes. Presumably other factors such as chromatin structure and cooperative binding play a role as well in determining the *in vivo* specificity of the associated transcription factors.

1. Schneider, T.D. & Stephens, R.M. Sequence Logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
2. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
3. Workman, C.T. *et al.* EnoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.* **33** (Web Server Issue), W389–W392 (2005).
4. Matys, V. *et al.* TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34** suppl. Database issue, D108–D110 (2006).
5. Vlieghe, D. *et al.* A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34** suppl. Database issue, D95–D97 (2006).
6. Teixeira, M.C. *et al.* The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **34** suppl. Database issue, D446-451 (2006).
7. Zhu, J. & Zhang, M.Q. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* **15**, 607–611 (1999).
8. Salgado, H. *et al.* RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.* **34** suppl. Database issue, D394–D397 (2006).
9. Munch, R. *et al.* PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.* **31**, 266–269 (2003).
10. Djordjevic, M., Sengupta, A.M. & Shraiman, B.I. A biophysical approach to transcription factor binding site discovery. *Genome Res.* **13**, 2381–2390 (2003).