What is (in) a sequence logo?

Given a collection of aligned binding sequences (i.e. binding sites) for a particular protein, the most common (and convenient in terms of accuracy and simplicity[1]) representation of the binding motif is a Position Specific Frequency Matrix (PSFM).

ATGACATCAT	ATTCGCTAAT	ATTGCGAGAT	GTGTGATCAT	ATGTTGCCAG
ATGCGACAAT	GCTAGCTCAG	ATGCTGATAT	GTACTGACAT	ATGAGATTAT
ATGCTGCCAA	TAGCTAGCAT	TTGTGATGAT	ATGCATTCAG	ATCAGACCAT
ATGCGATAGG	ATCGCGCCAT	TTAGCATGCC	ATGAATACTT	ATGACAGCAT
ATCGACGTAC	ATCGCTACAT	ATTGCATCAG	ATGGACCCCT	ATGATGACTT

Table 1 – List (or collection) of binding sites for the hypothetical protein BUH.

	1	2	3	4	5	6	7	8	9	10
A	0.76	0.04	0.08	0.28	0.12	0.44	0.24	0.12	0.80	0.04
C	0.00	0.04	0.12	0.32	0.28	0.12	0.28	0.68	0.08	0.04
T	0.12	0.92	0.16	0.16	0.28	0.12	0.40	0.08	0.08	0.68
G	0.12	0.00	0.64	0.24	0.32	0.32	0.08	0.12	0.04	0.24

Table 2 – Position Specific Frequency Matrix for transcription factor BUH.

Tables 1 and 2 show the list of sites, and corresponding PSFM, for the hypothetical protein BUH. The PSFM is obtained from the binding site list (i.e. the binding motif) by measuring the relative frequencies of each base at each position in the binding motif. For instance, the `0.76` frequency of A in position 1 comes from finding 19 A occurrences in the first position of the motif, among a total of 25 sequences (`19/25=0.76`).

It is quite intuitive that the PSFM represents the probability of finding a particular base at a particular position in the motif. What is not so intuitive is the idea that these probabilities can be seen as conditional probabilities in an information transmission process (see A gentle introduction to information content in transcription factor binding sites for an extensive discussion on this issue).

[1] One can consider many other representations that do not make the assumptions inherent to a PSFM, such as models taking into account interdependency between motif positions, but these models tend to be much more complex without (seemingly) a comparable return in terms of accuracy in describing the binding motif.

Basically, we can consider any given random sequence of length *L* (i.e. a random *site*) in the genome and compute the probability of observing any particular base at any position within that site. These probabilities are our best initial guess as to what base occupies each position and can be approximated by the background genome frequency for each base. Based on these probabilities, we can compute our original uncertainty, which we will call *H(X)*. To reduce our uncertainty to zero and be completely sure of what base occupies each position in our random site (i.e. gain information about the site sequence), we would need to board the *Proteus* submarine from Fantastic Voyage and take a good look at the particular stretch of DNA sequence we selected at random. Alternatively, we could go the NCBI's Genbank database and download the whole genome sequence (which we already used to compute the background frequencies), look at the chosen random position and be done with it.

But let us suppose that we live in a pre-genomic era or that (even worse!) the NCBI site is down! How could we gain information on the sequence of our random site? Well, one not very obvious way would be to start throwing proteins at it and see if any of them bind it. Observing binding of a particular protein, like BUH, to our randomly chosen site is a roundabout, yet still quite effective, way of gaining some information about the site sequence. The only nag is that observing BUH binding the site is not quite the same as getting there with *Proteus* and taking a look ourselves. Proteins can be quite specific in the sequence of the targets they bind, but there is always some amount of noise in the recognition process. We have already seen that. The probabilities in the PSFM are the probabilities of each base occupying each position of the site *once we know* that the protein binds there. Thus, they are conditional probabilities *P(X|Y)* and they express our remaining uncertainty over what base occupies each position (X) *after* observing binding of the protein (Y). If we take the difference between our original uncertainty *H(X)* and our new uncertainty *H(X|Y)*, what we get is the reduction in uncertainty we experience about the base occupying each position when we observe binding of a protein to a previously unknown sequence.

The average uncertainty over something is typically expressed by a measure called *entropy*[2], which can be defined as:

[2] Again, see A gentle introduction to information content in transcription factor binding sites for a more comprehensive introduction to this issue.

$$H(X) = -\sum_{i=1}^{N} \left(p_i \cdot \log(p_i)\right), \quad N = number\ of\ \ possible\ outcomes\ /\ states \qquad (1)$$

Entropy was introduced by Shannon (Shannon 1948) as a weighted average of uncertainty or, equivalently, information (as the more uncertain we are about something, the more information we gain by observing it)[3].

Since we know the genomic base frequencies (`p(X=A)`, `p(X=G)`…), we can compute the original entropy *H(X)* of any position of our site. And since from the PSFM we know the conditional probabilities (`P(X=A|Y=protein binding)`, `P(X=G|Y=protein binding)`), we can also compute the a posteriori entropy *H(X|Y)*. The difference between both entropies is called mutual information and is the main message conveyed by a sequence logo. Mutual information expresses the reduction in uncertainty (or information gain) over what base occupies each position in a given site we experience by knowing that a protein binds the site. Mutual information per position can be expressed as a line plot (

Figure 1), just like the frequency information of a PSFM can be expressed with a frequency logo (
Figure 2).

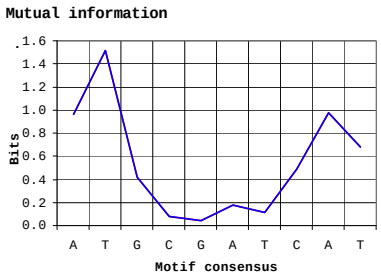

Figure 1 – Mutual information plot for the BUH binding motif using the collection in **Table 1** as input.

[3] Some people prefer to use the word uncertainty, rather than entropy, for the average uncertainty. This is because Shannon's entropy and physical entropy are only tangentially related, and the use of the "entropy" term for Shannon's information measure can often be confusing. For more about this, see Tillman, F. and B. Roswell Russell (1961). "Information and entropy." Synthese **13**(3): 233-241.

Figure 2 – Frequency logo for the BUH binding motif (generated through http://genome.tugraz.at/Logo/) using the collection in **Table 1** as input.

The beauty of sequence logos (
Figure 3) (Schneider and Stephens 1990) is that they combine both types of information in a single, intuitive, graphical representation.

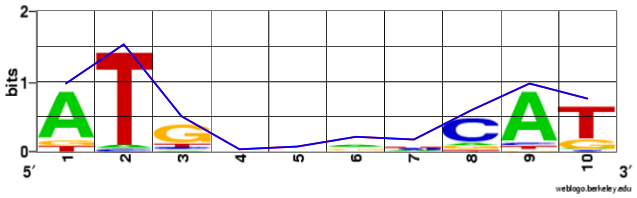

Figure 3 – Sequence logo for the binding motif BUH generated using the WebLogo server (Crooks, Hon et al. 2004). The mutual information function is superimposed on the logo.

Thus, the height of each letter is proportional to its frequency at that particular position, while the height of the stack indicates the mutual information in that position. Since sites have to be actively maintained by evolution if they are to remain functional, mutual information (i.e. stack height) is often interpreted as an indicator of evolutionary conservation at each position.

References

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: a sequence logo generator." Genome Res **14**(6): 1188-90.

Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Res **18**(20): 6097-100.

Shannon, C. E. (1948). "A mathematical theory of communication." Bell System Technical Journal **27**: 379-423 623-656.

Tillman, F. and B. Roswell Russell (1961). "Information and entropy." Synthese **13**(3): 233-241.