DNA Binding Proteins

• Job: to regulate transcription of DNA

• The DNA
  – Phosphate backbone is uniform for the whole DNA molecule but there is a major
    and minor groove
  – Different chemical environment in the grooves depending on the base pairs at a
    particular site
  – base pairs especially exposed in the major groove

• Example: helix-turn-helix binding motif
  – alpha helix then a turn then an alpha helix – forms a tertiary structure that fits
    well in the major groove
  – motif is seen frequently among DNA binding proteins
  – often form dimers that can bind at multiple points on the helix – a single molecule
    cannot target specifically because it can only be unique for a small number of base
    pairs (such as just “AAGA”)

• Example: zinc finger motif
  – zinc ion holds together a beta sheet and alpha helix which fits into the major
    groove
  – amino acids projecting from the alpha helix interact with the base pairs

• Example: leucine zipper motif
  – homo-dimer – two identical proteins that form one functional dimer
  – two long (separate) alpha helices with lots of leucine
  – can mix and match alpha helices to make heter-dimers with different affinities

• often interactions are formed via hydrogen bonding with the base pairs – all base pairs
  of protruding alcohol or amine groups that can easily form H-bonds

• it is difficult to impossible to “predict” the DNA binding code because interactions
  are not always straightforward – can interact with various strands, and can assume
  any number of positions along the DNA; additionally, the protein may undergo a
  conformational change when binding the protein and may bend the DNA itself. Some
  proteins interact with pieces of DNA far away from each other on the strand
• Example: bacterial met repressor
  – must be activated by SAM (S-adenosyl methionine, an important metabolite derived from methionine) – causes a conformational change
  – once activated, binds to the DNA and represses transcription of methionine synthesis genes
  – Thus, cell can regulate methionine/SAM production by downregulating its production when the pathway’s end product is more common
  – Everything is ephemeral – all proteins eventually are degraded, can fall off or reattach to the DNA; SAM molecules can attach and fall off the protein, etc. All of biology is based on equilibrium conditions changing and many small events being more or less probable.

• Example: the TATA box
  – found in *E. coli* and other bacteria (especially Eubacteria)
  – always found about 10 basepairs upstream from a transcription start
  – Consensus sequence – TATAAT is the expected sequence, but approximate matches are allowed because of general affinity of proteins for the nucleotides
  – e.g. might be able to tolerate either pyrimidine (A or G) because they are structurally similar to each other (both concentric rings)
  – the farther away from TATAAT you get, the less likely it is that any given transcription protein molecule will bind there – it is still possible and given the number of proteins floating around, it will still happen, but not as often
  – almost no perfect matches – how do we identify instances? how do we identify consensus sequences?
  – statistically make a table of frequencies of seeing a given letter in a given position of a TATA box
  – change these frequencies into scores (positive and negative) and sum over the particular scores for a particular sequence – can get a score for every position on the genome
  – if you draw 6 random letters and score them according to the table, scores will have some mean (and be very roughly normally distributed); if you draw from the probability distribution implied by the table, however, you’ll get a higher mean.
  – statistics
    * what is the probability of getting a sequence *S* if we assume that it arises from the TATA box: \( P(S|\text{“tata”}) \)
    * What is the probability of getting a sequence *S* if we don’t assume the TATA box: \( P(S|\text{“non-tata”}) \)
    * these can both be calculated fairly easily using basic frequencies
    * the log of the ratio of these is used as the score
a score might be \(-\infty\); perhaps replace with some low value like \(-46\)? more below

the frequency counts per position is the maximum likelihood estimator for the model (this is intuitive)

Complication – DNA isn’t always evenly distributed in terms of base pairs, thus the frequencies observed might actually be due to just normal DNA sequences – this is why we have the denominator of the likelihood

Information Theory

\begin{itemize}
  \item Relative entropy – how much information is shared or not shared between two variables
  \item entropy can be thought of as the amount of “space” necessary to store something
  \item \(H(P||Q)\) – relative entropy of \(P\) to \(Q\)
  \item if the information in \(P(x)\) exactly predicts the information in \(Q(x)\), then the relative entropy is 0 (the distributions are identical)
  \item \(H(P||Q)\) is the expected score from the model with \(P(x)\) as the prob. of sequence \(x\) and \(Q(x)\) as the background prob. of sequence \(x\)
  \item \(-H(Q||P)\) is the expected score from the DNA overall
\end{itemize}

Can use a pseudocount to prevent the \(-\infty\) values – e.g. add 0.5 to each count before finding log scores

Given unaligned sequences thought to contain some motif, how do we find it?

– might have a set of upstream regions of known genes but not know what sequences regulate those genes

– Idea: look for maximum relative entropy between the sequences. Unfortunately, this is NP-hard; best we can afford to do (probably) is approximate it some how. Three approaches to be presented

– Greedy approach

\begin{itemize}
  \item \(k\) sequences, \(s_1, s_2, \ldots, s_k\)
  \item motif length is \(l\); breadth is \(d\)
  \item start exhaustively enumerating subsets of length \(l\) subsequences
  \item compute relative entropy of each subset and throw away all but \(d\) best
  \item the larger \(d\), the better it runs, but the slower it runs
\end{itemize}

– Expectation Maximization approach – like a hidden Markov model approach – we have hidden data (motifs) and visible data (sequences), so try to find the motifs that have the maximal expectation; iterate and home in on the right sequences. Details next time.