20.320, notes for 12/6

Thursday, December 06, 2012
9:41 AM

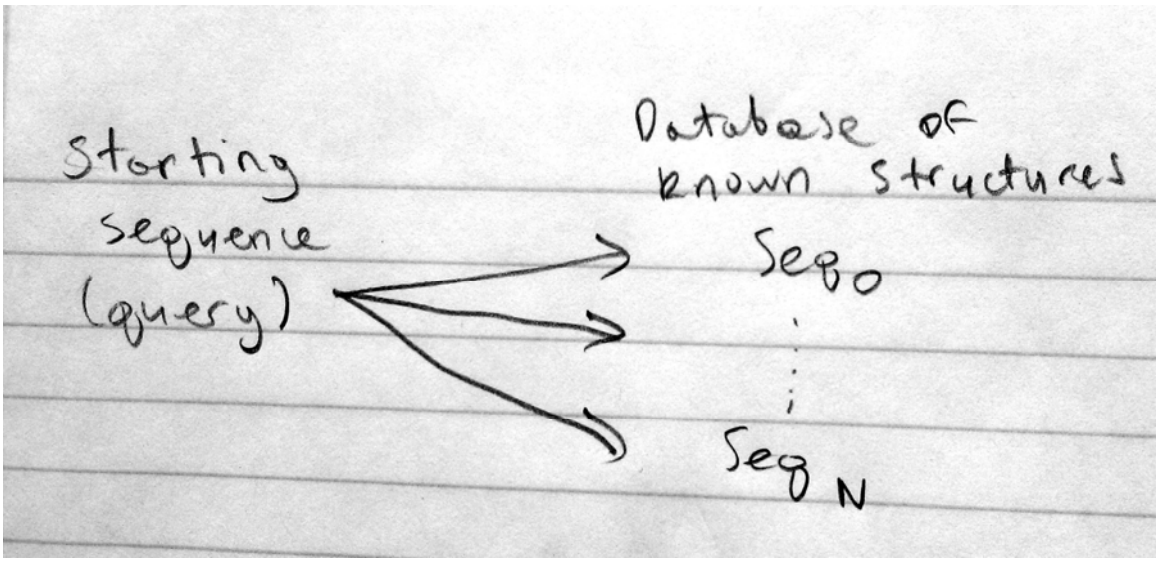
Structure from Sequence

So, given all the predictions that we've made so far, to what extent can we predict protein structure? We're currently able to do only small proteins, and with an immense cost of computational resources. Can we do better?

CASP, Critical Assessment of Protein Structure, is a competition where the unbiased predictions of structural biologists are used to determine which of the computational techniques work best. This has yielded some very important insights. For starters, it's a bad idea to try to predict structure from entirely first principles. The best you can do is find similar proteins that have already been crystallographically determined, and use that as a starting point. We've talked about how to refine a final structure, but today we'll talk about sequence homology and how to get from there to a starting structure.

Let's start at the beginning. We'll continue towards the end, and when we get there, we will stop.

We start with a sequence query, which we can compare to a database of known protein sequences. How do we do this? One really useful thing is to classify the database by the protein domains that it contains. Remember that many different proteins share very similar domains. These domains are regions that perform the same task and have similar form. However, these domains are usually located in different places on the protein sequence. Furthermore, there might be lots of variation between functionally similar sequences. It's crucial to get a probabilistic measure of whether a candidate protein matches your candidate. That will be the main content of this lecture.

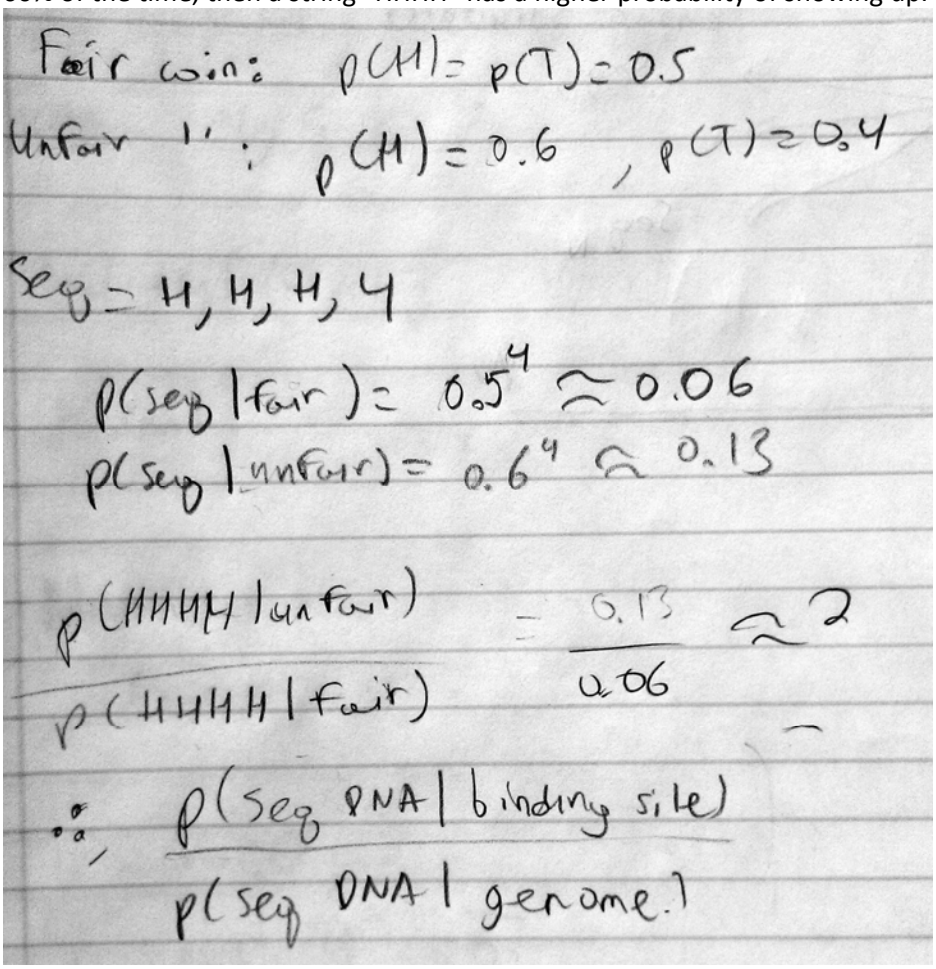


One way to do this is to go through the positions in the protein sequence and determining the probability of that amino acid being there. The more conserved any amino acid is, the higher the probability of finding it in its position. Here we can draw some insight from DNA analysis, where we similarly try to find homology between sequences. We may know many wild type sequences of a given element (say, a txn factor binding site). We can compare those and count the occurrence of A, G, C, and T in the population. From that we determine the probability of each base occurring there. If we assume that all positions are independent, then the probability of the whole sequence is the product of that of the individual positions.

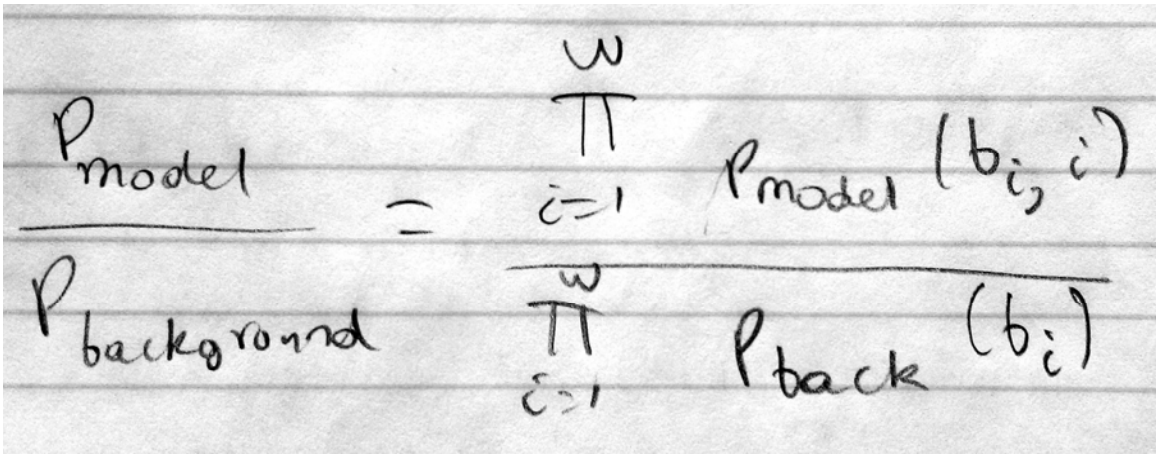
$$P(sequence) = \prod_{i=1}^{N} p(X_i)$$

In reality, though, we have some hidden assumptions. The above is actually the probability ***given*** that what you have is the sort of structure you're looking for (in our example, a binding site). The only other possibility here is that the structure is a binding site for something else. Thus, this is not the best way to approach our problem in general.

We can draw a useful analogy from coin-tossing probabilities. If we have a n unfair coin that lands heads 60% of the time, then a string "HHHH" has a higher probability of showing up.

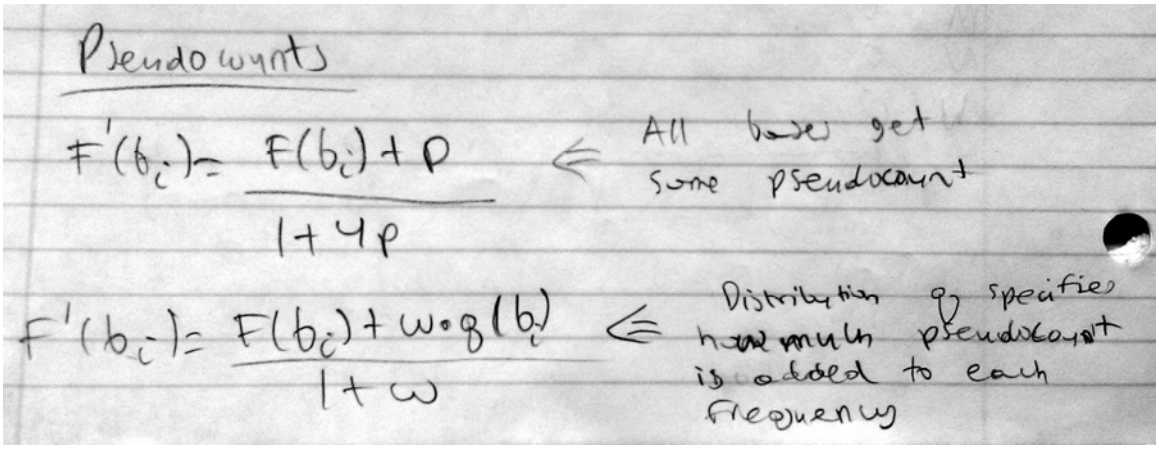


The useful number for our actual example is a ratio of probabilities, the probability of a sequence being the result of our model over the probability that it's a result of our background (that it just happens a lot in the genome).



There's a problem with our model, though. If something happens in a query sequence that *never* happened in the sequences building the model, the total probability assigned to that query will be zero. We know there's an energetic cost, but surely the probability isn't *zero*, especially if we built our model on a small number of examples. So then what? Well, there's this thing called pseudocounts.

We want to avoid categorical statements when dealing with small model-building datasets. The idea is to assign a small baseline probability to every option, even if it never happened in the model. All probabilities are slightly boosted, and now nothing is impossible.



We should add, by the way, that we're usually dealing with very small numbers and ratios thereof. Therefore, the convention in the field is to use not probabilities, but logarithms (base 2) of probabilities. All the techniques described here for DNA work the same way for proteins. There's websites, like Prosite, that have already computed the sequence probabilities for a whole bunch of known domains. Thus, you can query it for any new sequence that you're trying to analyze.

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012