Lecture 13

Markov Chains

February 17, 2000
Notes: Jonathan Schaefer

In Lecture 11 we discovered that correlations between sequence positions are significant, and should often be taken into account. In particular, in Section 11.4 we noted that codons displayed a significant bias, and that this could be used as a basis for finding coding regions. Lecture 12 then explored algorithms for doing exactly that.

In some sense, Lecture 12 regressed from the lesson of Lecture 11. Although it was using codon bias to score codons, it did not exploit the possible correlation between adjacent codons. Even worse, each codon was scored independently and the scores added, so that the codon score does not even depend on the position the codon occupies.

This lecture rectifies these shortcomings by taking codon correlation into account in predicting coding regions. In order to do so, we first introduce Markov chains as a model of correlation.

13.1. Introduction to Markov Chains

The end of Section 11.2 mentioned "...a random sequence ...generated by a process according to dinucleotide distribution P", without giving any indication of what such a random process might look like. Such a random process is called a "Markov chain", and is more complex than a process that draws successive elements independently from a probability distribution. The definition of Markov chain will actually generalize dinucleotide dependence to the case in which the identity of the current residue depends on the previous k residues, rather than just the previous one.

Definition 13.1: Let S be a set of states (e.g., $S = \{A, C, G, T\}$). Let $(X_0, X_1, X_2, \ldots)$ be a sequence of random variables, each with sample space S. A *kth order Markov chain* satisfies

$$\Pr(X_t = x \mid X_0, X_1, \ldots, X_{t-1}) = \Pr(X_t = x \mid X_{t-k}, X_{t-k+1}, \ldots, X_{t-1})$$

for any t and any $x \in S$.

In words, in a kth order Markov chain, the distribution of X_t depends only on the k variables immediately preceding it. In a 1st order Markov chain, for example, the distribution of X_t depends only on X_{t-1}. Thus, a 1st order Markov chain models diresidue dependencies, as discussed in Section 11.2. A 0th order Markov chain is just the familiar independence model, where X_t does not depend on any other variables.

Markov chains are not restricted to modeling positional dependencies in sequences. In fact, the more usual applications are to time dependencies, as in the following illustrative example.

Example 13.2: This example is called "a random walk on the infinite 2-dimensional grid". Imagine an infinite grid of streets and intersections, where all the streets run either east-west or north-south. Suppose you are trying to find a friend who is standing at one specific intersection, but you are lost and all street signs are missing. You decide to use the following algorithm: if your friend is not standing at your current intersection, choose one of the four directions (N, E, S, or W) randomly and uniformly, and walk one block in that direction. Repeat until you find your friend.

This is an example of a 1st order Markov chain, where each intersection is a state, and X_t is the intersection where you stand after t steps. Notice that the distribution of X_t depends only on the value of X_{t-1}, and is completely independent of the path that you took to arrive at X_{t-1}.

Definition 13.3: A kth order Markov chain is said to be *stationary* if, for all t and u,

$$\Pr(X_t = x \mid X_{t-k}, X_{t-k+1}, \ldots, X_{t-1}) = \Pr(X_u = x \mid X_{u-k}, X_{u-k+1}, \ldots, X_{u-1}).$$

That is, in a stationary Markov chain, the distribution of X_t is independent of the value of t, and depends only on the previous k variables. The random walk of Example 13.2 is an example of a stationary 1st order Markov chain.

13.2. Biological Application of Markov Chains

Markov chains can be used to model biological sequences. We will assume a directional dependence and always work in one direction, for example, from $5'$ to $3'$, or N-terminal to C-terminal.

Given a sequence s, and given a Markov chain M, a basic question to answer is, "What is the probability that the sequence s was generated by the Markov chain M?" For instance, if we were modeling diresidue dependencies with a 1st order Markov chain M, we would need to be able to determine what probabilities M assigns to various sequences.

Consider, for simplicity, a stationary 1st order Markov chain M. Let $A_{r,s} = \Pr(X_t = s \mid X_{t-1} = r)$. A is called the *probability transition matrix* for M. The dimensions of the matrix A are $|S| \times |S|$, where S is the state space. For nucleotide sequences, for example, A is 4×4.

Then the probability that the sequence $s = (s_0, s_1, \ldots, s_t)$ was generated by M is

$$\Pr(s_0 s_1 \ldots s_t) = \Pr(s_0) A_{s_0,s_1} A_{s_1,s_2} \cdots A_{s_{t-1},s_t} = \Pr(s_0) \prod_{i=1}^{t} A_{s_{i-1},s_i}.$$

In this equation,

1. $\Pr(s_0)$ is estimated by the frequency of s_0 in the genome, and
2. $A_{r,s}$ is estimated by $N_{r,s}/N_r$, where $N_{r,s}$ is the number of occurrences in the genome of the diresidue (r, s), and $N_r = \sum_i N_{r,i}$.

Markov chains have some weaknesses as models of biological sequences:

1. Unidirectionality: the residue s_i is equally dependent on both s_{i-1} and s_{i+1}, yet the Markov chain only models its dependence on the k residues on one side of s_i.

2. Mononucleotide repeats are not adequately modeled. They are much more frequent in biological sequences than predicted by a Markov chain. This frequency is likely due to DNA polymerase slippage during replication, as discussed in Example 11.2.

3. Codon position biases (as discussed in Section 11.4) are not accurately modeled.

13.3. Using Markov Chains to Find Genes

We will consider two gene finding algorithms, GeneMark [1, 2] and Glimmer [3, 4]. Both are commonly used to find intron-free protein-coding regions (usually in prokaryotes), and both are based on the ideas of Markov chains. As in Section 12.1, both assume that a training set of coding regions is available, but unlike that method, the training set is used to train a Markov chain.

GeneMark [1, 2] uses a kth order Markov chain to find coding regions, where $k = 5$. This choice allows any residue to depend on all the residues in its codon and the immediately preceding codon.

The training set consists of coding sequences identified by either long open reading frames or high sequence similarity to know genes.

Three separate Markov chains are constructed from the training set, one for each of the three possible positions in the reading frame. For any one of these reading frame positions, the Markov chain is built by tabulating the frequencies of all $(k + 1)$-mers (that is, all length $k + 1$ substrings) that end in that reading frame position. These three Markov chains are then alternated to form a single nonstationary kth order Markov chain M that models the training set.

Given a candidate ORF x, we can compute the probability p that x was generated by M, as described in Section 13.2. This ORF will be selected for further consideration if p is above some predetermined threshold. The "further consideration" will deal with possible pairwise overlaps of such selected ORFs, in a way to be described in the next lecture.

References

[1] M. Borodovsky and J. McIninch. GeneMark: Parallel gene recognition for both DNA strands. *Comp. Chem.*, 17(2):123–132, 1993.

[2] M. Borodovsky, J. McIninch, E. Koonin, K. Rudd, C. Medigue, and A. Danchin. Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Research*, 23(17):3554–3562, 1995.

[3] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641, 1999.

[4] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2):544–548, 1998.