Monday, September 10, 2012

0.1 Dynamic programming and DPP

For concreteness, we focus on Nussinov's DP algorithm. The ideas should work for general DPs. Suppose we have a fixed DNA/RNA sequence S of length L. Let y denote a particular secondary folding structure of the sequence. We try to follow the notations of the Supplementary Methods of Taskar's SDPP paper.
The factor graph has $T = L(L-1)$ number of variable nodes, indexed by $t = (i, j), i < j$. The possible values for the variable node (i, j) is the set of energies possible for the subsequence $S(i, j)$. There are also $L(L-1)$ factor nodes indexed by $(i, j), i < j$. The factor node $F(i, j)$ is connected to the variable nodes $\{(i+1, j-1), (i+k, j), (k+1, j)\}, i < k < j$. For $(i, j) \neq (1, L)$, the factor $F(i, j) = 1$ if the substructure $y(i, j)$ has a consistent folding (no mismatched base-pairs, etc), and $F(i, j) = 0$ if the substructure is not consistent. For the "last" factor, $F(1, L) =$ energy of the structure y.

Saturday, September 22, 2012

To fill in a bit more and correct some mistakes from above.
We have a factor graph with $T = \frac{L(L-1)}{2}$ variable nodes, indexed by $t = (i, j)$. Similarly, there are $\frac{L(L-1)}{2}$ factor nodes that we denote by $F(i, j)$. The factor node $F(i, j)$ is connected to the variable nodes $\{(i+1, j-1), (k, j), (i, k)\}, i < k < j$. Generic variable nodes are denoted by t, and generic factor notes are denoted by α.

1. The values for y_t are the possible folding scores for the subsequence $S(i, j)$. In the simplest case, the folding score equals to the number of base pairings, and $y_t \in \{0, 1, ..., L/2\}$.

2. For a factor node α, we think of y_α as a list of values $\{y_{t_1}, y_{t_2}, ...\}$ where t_i is a variable node connected to α. Let $\alpha*$ denote the final factor node $F(1, L)$.

3. A factor node is associated with weight $w_\alpha(y_\alpha)$. We have

$$w_\alpha(y_\alpha) = (q^2(y_\alpha), q^2(y_\alpha)\phi_r(y_\alpha), q^2(y_\alpha)\phi_l(y_\alpha), q^2(y_\alpha)\phi_r(y_\alpha)\phi_l(y_\alpha))$$

4. For $\alpha \neq \alpha*$, $q(y_\alpha) = 1$ if the list y_α is feasible, i.e. the score y_{ij} can be feasibly arrived from $y_{(i+1)(j-1)}, y_{kj}, y_{ik}$ and the sequence $S(i, j)$. And $q(y_\alpha) = 0$ otherwise. For $\alpha*$, $y_{\alpha*} = y_{1L}$, the score of the entire structure.

Questions to address:

1. The two pass belief propagation computes $\sum_y \prod_\alpha w_\alpha(y_\alpha)$, where $y = \{y_{ij}\}$ is a list of scores over all substructures. What we actually need to compute is $\sum_s \prod_\alpha w_\alpha(y_s)$, where the sum is over all feasible substructures. Seems like there is an straightforward bijection between y and the set of all possible structures for S. Need to verify.

2. For the two pass belief propagation to work, i.e. equation 3 of the supplement, we need the factor graph to be a tree. This is not true in our case, which has many loops. However, the weights at the factor nodes are simple, $w_\alpha(y_\alpha = (0, 0, 0, 0))$ if y_α is not feasible. So might still work.

3. As we discussed, the factor graph is densely connected. But for a given structure, only a few variable nodes y_t contribute. Need to work this out precisely.