Reading Assignment 2

Najib Ishaq

In this paper the authors pioneer a method to consider long-range beta-strand dependencies for homology-detection between well-annotated protein structures. The authors point out that previously existing HMMs (Hidden Markov Models) are good at detecting homologs from local residue interactions but that they are poor at capturing long-range dependencies. Even the generalization to MRFs (Markov Random Fields), while increasing predictive power, suffer from the same exponentially increasing computational complexity.

The authors present a suite of new methods, MRFy, which build upon existing Markov models by adding non-local hydrogen-bonded beta-strands as additional edges in the graph. Insertions and deletions in these strands is prohibited and a consensus (of at least half the participating sequences) is used to reduce complexity and improve predictions. The authors modify the original Viterbi recurrence relation with conditional probabilities representing the new edges. These relations are then solved after transforming them to negative log space.

The authors then argue for the necessity for stochastic search by proving the exponential complexity of MRFs. Stochastic search, in one of the strategies as implemented in MRFy, proceeds in three steps:

Make an initial guess using one of five options:

1. A *random-placement-model* that guesses beta-strand placements under the constraint that no singe placement of a beta-strand invalidates the placement of any other beta-strand.
2. A *secondary-structure* prediction from PSIPRED used to judge beta-strand placements.
3. A *template-based model* based on the fact that homologs should have roughly similar beta-strand placements.
4. A *multi-scaled template-based model* that uses multiple local-homologs as templates as opposed to a single best global homolog.
5. A hybrid of SMURFLite and MRFy. A given template is used to generate a new one by first removing and beta-strands with a high inter-leave threshold and then greedily adding them back by moving existing beta-strands. The result is used as the initial guess.

Optimize the augmented Viterbi score for the guess using one of three options:

1. *Simulated annealing*, a heuristic for stochastic search that uses an *acceptance probability function* that randomly moves to improved states but tends to revert to traditional hill-climbing as time progresses. This can be

augmented by starting with multiple different guesses, processing them in parallel, and finally using the best scoring structure.

2. A *Genetic algorithm* that mimics evolution via natural selection and semi-random mutations. For each generation some number of the fittest individuals are selected to "reproduce" into the next generation. Fitness is assessed by the Viterbi score. Sexual reproduction is simulated by randomly pairing up the selected individuals. Each pair produces some offspring that randomly mutate some aspects of the parents' characteristics. The random mutations are constrained in that beta-strands are not allowed to shift too far based on their neighbors.

3. *Local search* that explores the immediate neighborhood of a structure in an attempt to reach a local optimum. After reaching a local optimum, the population is "diversified" by a *non-local* move. Local search is then repeated.

Decide when to terminate the optimization process using one, or a combination of, of three options:

1. The model can be stopped after a user-specified number of generations have been processed.

2. The model can be stopped after a user-specified amount of time has elapsed.

3. The model can be stopped after it has converged, i.e. there has been no improvement over a user-specified number of generations.

MRFy offers offers several significantly different strategies with a number of tunable hyper-parameters. The authors performed a search over these parameters using a small dataset and compared the results to state-of-the-art methods. The search is easily parallelized because the implementation is in Haskell. Remote homology detection accuracy was improved by MRFy to an AUCROC of 0.95. Alignment quality was on-par with, and often better than, that of SMURFLite.