Midterm 1 Review

Here’s a list of things we’ve covered so far, and a set of sample questions. These are questions that failed to make it onto the midterm for one reason or another. Usually that means they are somewhat ambiguous or too specific or of questionable relevance, though some of them were left off just out of time considerations. As such, you should use these questions as an upper bound on the level of sophistication you will need for the midterm, though in practice the midterm questions will tend to be slightly easier than the hardest questions in this list (though I suppose that’s a matter of opinion…).

You should also go over all the homework problems and make sure you understand the concepts behind them. Generally, you should not need to worry about equations, numbers, etc, except you should know how to do a dynamic programming alignment using Needleman-Wunsch and Smith-Waterman with fixed gap penalties, read off and use values from a substitution matrix, and calculate a score for a motif given a positional weight matrix.

A rough guide to what we think you should know is:

1. Biology:
   * What’s a gene, what a gene looks like (exons, etc) how transcription, translation work, codons, coding regions vs noncoding regions, reading frame

2. Sequence Alignment:
   * Greedy algorithms, Dynamic programming algorithms
   * Needleman-Wunsch Global alignment; Smith-Waterman Local alignment (how to fill in a matrix and interpret the results)
   * Fixed gap penalty (be able to use); Affine gap penalty (theory of only)
   * PAM and BLOSUM matrices – how to use, where they come from

3. BLAST:
   * How the algorithm works – seeds, extending HSPs, etc.
   * Parameters – scoring matrix, low-complexity mask,
   * Understand what the basic types of BLAST do (blastn, blastp, blastx, PSI-BLAST.)
   * BLAST statistics – the concept of the random walk, the test statistic Ymax, significance testing and the hypotheses you’re testing (no need to worry about formulas or equations)
   * BLAST High score; bit score; expectation value; P-value.

4. Motif finding:
   * Splice sites; Acceptor; Donor; Intron; Exon.
   * PWMs – how to create (logged, background normalized variations), how to use, how to evaluate (information content, relative entropy)
   * Three algorithms for motif finding: Consensus; Gibbs Sampling (and how AlignACE modifies the basic algorithm); EM (and how MEME modifies the basic algorithm).
5. Promoter Prediction:
* the signals and features used to predict promoters (especially TSS; TATA box; CpG islands; DNA structure; first exon)
* Measures for evaluating how well your algorithm performs (sensitivity, specificity)

6. Comparative Genomics:
* Genome scale alignment strategies (BLASTZ; AVID; LAGAN)
* Reasons we care about genome-scale alignment / conservation, and how it is applied to various problems
* VISTA, rVISTA

7. Gene Finding:
* Challenges to finding genes, structure of a gene (UTRs, exons, introns, stop codon, etc)
* Identifying genes based on analyses of open reading frames

Fun Questions

1. Describe what a greedy algorithm is, and give one example of a greedy algorithm we've covered so far, briefly explaining why it would be considered a greedy algorithm.

2. Explain the sensitivity / specificity trade-off, using one algorithm that we’ve seen or used so far as an example.

3. In Psi-BLAST, after each iteration, a positional weight matrix is created based on the current set of highly significant sequences, and this PWM is used to do another round of BLAST. Another way to do the search would be to run a separate BLAST on each sequence in the set and return all significant matches over each separate BLAST. Compare these two approaches by describing what sorts of database matches you would expect each approach would yield after several iterations.

4. Let's take a substitution matrix such as BLOSUM50 and create a new substitution matrix by doubling each of the entries in the original matrix -- positive scores will be twice as big, negative scores will be twice as negative. Call the new matrix BLOSUM50x2. Now use each of them to perform the same BLAST alignment search. Which (if any) of the two matrices will report hits that have a lower e score? Explain. If you feel that your answer depends on other factors, list what those are.

5. Given the following DNA sequence ATGGATTTAGGT, clearly label the six possible reading frames.

6. Explain why BLAST would be a poor tool to search for motifs.
7. How does knowing where the first exon of a gene help in identifying that gene's promoter? Would it be more or less useful to know where the start codon is? Why?

8. Neither the Gibbs sampling nor the Expectation-Maximization strategies for motif finding are guaranteed to find the motif that you’re actually interested in. Choose either the Gibbs or EM strategy and describe a way your chosen algorithm might converge on a sub-optimal motif. Describe one way in which you can reduce the risk of this happening.

9. Describe how rVISTA uses sequence conservation information for promoter prediction. Explain how the conservation information can be useful for this task, and list two situations where you might not want to use rVISTA for promoter prediction.

10. Why would a full dynamic-programming approach not work for most genome alignment tasks? Would using BLAST be useful instead?

11. In random walk theory as applied to BLAST scores, what is meant by a “ladder point” and how are these important in calculating the BLAST score?