Suggestions for group projects (some publishable).
Additional information will be given once you express interest.

1. **Sequence search.** Biologists frequently like to find the sequences which are most similar to a sequence of interest. There are currently so many sequences that it is intractable to exhaustively search for similarities. Therefore, heuristic methods have been developed for doing this (especially, BLAST). Download a copy of the SwissProt data set (about 1/2 gig), and develop your own *tractable* algorithm for finding sequence similarity. The algorithm isn't expected to be good, but it should at least be better than random. The algorithm is not expected to process all data on-the-fly. You can preprocess the data and store it for later retrieval by your algorithm. For example, the BLAST algorithm has broken up sequences into $k$-mers; i.e., it has stored which sequences match any given sequence of $k$ amino acids (I think $k = 5$).

2. **Sequence alignment.** Dynamic programming is the method of choice for sequence alignment. However, other methods are invented from time to time. Invent your own algorithm for pairwise or multiple sequence alignment. Examples:
   - Start with a dotplot
   - Use a genetic algorithm

3. **Multiple sequence alignment.** Find or create a heuristic that has not been used for multiple sequence alignment and see how it fares.

4. **Predicting protein function.** Proteins that share a function are often located near one another in gene and protein networks. Use combinatorial analysis, including mutual clustering coefficient and other cumulative hypergeometric-based methods, to improve prediction of proteins that belong to a function or process (e.g., DNA synthesis or embryo development).

5. **Clustering gene expression data:** Clustering algorithms suffer from having to group genes across all samples, or group samples across all genes. Biclustering is a form of clustering which simultaneously clusters subsets of genes across subsets of samples. There are many approaches to biclustering because the problem is NP Hard. Find a subspace clustering algorithm that has not been used with biological data and see how it fares.

6. **Clustering biological networks:** Biological networks are often clustered or partitioned to predict function for uncharacterized proteins and/or to predict protein complexes and molecular machines. Find or create a graph clustering (or
partitioning) algorithm that has not been used with biological data and see how it fares.

7. **Information measures for protein alignments.** Shannon information, the decrease in uncertainty (where uncertainty is measured as the sum of the weighted log probabilities), is often used to characterize the amount of information in a position or range of a sequence alignment. In particular, it is often used to characterize motifs. However, this information measure assumes that all symbols are equally different, whereas some amino acids (e.g. leucine, isoleucine and valine) are much more similar to each other than to other amino acids. Is there a way to characterize the information content of a sequence alignment taking these similarities into account? Alternatively, is there some measure other than information that could be used in this context?

8. **Comparison of methods for estimating the probability of modular/correlated motifs in sequences.** We have recently developed new methods based on state machines and transition matrices to calculate the probability of modular and/or correlated motifs in sequences. Several earlier methods for estimating these probabilities are available. How does each of these methods perform on biological motifs of different kinds, such as transcription factor binding sites, ribozymes, and splice sites?

9. **Is there an intrinsic length scale for RNA folding?** RNA secondary structure is defined by base pairing. However, statistical considerations suggest that complementary halves of a helix that are very distant in the sequence are unlikely to pair with each other because "good enough" choices are likely to occur in the sequence separating them. In a random sequence that contains one instance of one helix half, and two instances of the other helix half, how does the propensity to pair with each helix depend on the distance between the halves? How do the effects scale with "distractors" of different quality, or different numbers of distractors? The overall goal is to use information from known RNA structures to improve prior probabilities for structure prediction.

10. **Identification of protein domains.** Proteins are often built up out of modular pieces with particular functions, like binding to DNA or catalyzing a chemical reaction. Often, a particular domain will appear in many different proteins. Devise a tractable scheme for identifying a set of protein domains, using a new technique or heuristic, and see how well it works.

11. **DNA Computing.** Some problems that are difficult to solve, like the Traveling Salesman, may be more tractable to deal with in parallel. Some researchers, like Len Adelman (see http://www.popularmechanics.com/science/research/1281691.html? page=3) and Erik Winfree (http://www.dna.caltech.edu/Papers/thesis.pdf), have proposed using interactions between complementary DNA strands to compute fast solutions to problems like this. Describe such a computing system, and evaluate it. What advantages do you see in this design? What drawbacks might it have? Do you see a future for this sort of “wetware”?