SUGGESTED 206B CLASS PROJECTS (winter 2005)

1. Evolution and assembly of an extremely scrambled gene.

   Landweber et al., 2000 described the process of gene unscrambling. Gene unscrambling in ciliates represents one of nature's ingenious solutions to the problem of gene assembly. With some essential genes scrambled in as many as 51 pieces, these ciliates rebuild their fragmented genes and genomes. Landweber et al., 2000 report the complex pattern of scrambling in the DNA polymerase gene. The micronuclear copy of this gene is broken into 48 pieces. Direct repeats present at the boundaries between coding and noncoding sequences provide pointers to help guide assembly of the functional (macronuclear) gene. Landweber et al., 2000 investigated the evolution of this complex gene but their analysis involved no serious algorithmic study.

   PROBLEM: Formulate the gene unscrambling problem and solve it.

   REQUIREMENT: Ability to formulate problems and algorithmic skills. This project requires strong background in both algorithms and biology.

   ASSIGNED PAPERS: L. F. Landweber, T.C. Kuo, and E. A. Curtis. Proc. Natl. Acad. Sci. USA, 97, 3298-3303, 2000

   P.A. Pevzner. Computational Molecular Biology: An Algorithmic Approach. (2000) The MIT Press. Chapter 10.

2. Micro-rearrangements in mammalian genomes.

   Recent sequencing projects revealed a wide-spread phenomenon of micro-rearrangements in mammalian genomes. However, there are still very studies of micro-rearrangements and the question how micro-rearrangements affect the genomic architectures remains open. In particularly, it remains unclear whether the *breakpoint re-use effect* happens at the level of micro-rearrangements.

   PROBLEM: Study micro-rearrangements in mammalian genomes and answer the question whether micro-rearrangements tend to re-use breakpoints.

   REQUIREMENT: Algorithmic/programming skills

   ASSIGNED PAPERS: Pevzner P, Tesler G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res. 2003 Jan;13(1):37-45.

   D. Sankoff, P. Trinh. Chromosomal breakpoint re-use in the inference of genome sequence rearrangement. RECOMB 2004

3. Approximation algorithms for the Maximum Subgraph with Large Girth Problem and analysis of repeats in DNA sequences.

   The recently proposed RepeatGluer algorithm is based on the solution of the Maximum Subgraph with Large Girth (MSLG) Problem. The MSLG Problem aims to remove bulges from the (weighted) graphs arising in repeat analysis. It amounts to finding a maximum weight subgraph in the graph (i.e., a collection of edges of maximum total weight) that does not contain short cycles (cycles of length less than a parameter *girth*). The MSLG Problem with parameter $girth = \infty$ is the well-known and easy to solve Maximum Spanning Tree Problem. However, for an arbitrary *girth* the MSLG problem is far from easy and no approximation algorithms with provable performance are known for this problem. Since the graphs arising in repeat analysis are very large (millions and even billions of vertices), speed should be a key factor in your algorithmic design.

   PROBLEM: Devise an approximation algorithm for solving the Maximum Subgraph with Large Girth Problem.

REQUIREMENT: Strong algorithmic skills (approximation algorithms and combinatorial optimization).

ASSIGNED PAPERS: P.A. Pevzner, H. Tang and G. Tesler (2004) De novo repeat classification and fragment assembly. Eight International Conference on Computational Molecular Biology RECOMB 2004, San Diego, March 2004

4. Multiple Alignment of Proteins with Shuffled Domains.

   The popular multiple alignment algorithms (like CLUSTAL) have some shortcomings. Zhang and Waterman, 2003 recently developed a new algorithm for *global* multiple alignment of *DNA sequences* that is based on the graph-theoretical Eulerian path approach. While their algorithm is accurate for DNA sequences it is not clear how to generalize it for aligning of *protein sequences*, particularly proteins with shuffled domains. In another development Raphael et al., 2004 described a new approach for aligning proteins with shuffled domains. While this technique is accurate in defining domain boundaries, it is not clear whether it can accurately align the same domain across many proteins.

   PROBLEM: Improve the Raphael et al., 2004 approach for global multiple alignment of proteins and modify it for constructing the accurate representation of each domain. Benchmark your approach against other available multiple alignment tools. In addition you will have to develop a new visualization tool for representing multiple alignments of proteins with shuffled domains.

   REQUIREMENT: Algorithmic and implementation skills.

   ASSIGNED PAPERS: Y. Zhang and M.S. Waterman. J Comput Biol. 2003;10(6):803-19. An Eulerian path approach to global multiple alignment for DNA sequences.

   Raphael et al. (2004) Multiple Alignment of proteins with shuffled and repeated domains (submitted).

5. Yet Another Approach to *de novo* Repeat Classification

   One way to address the repeat classification problem is to construct a list of all frequent l-mers ($l = 14 - 20$) from a genome in a hope that such l-mers (if assembled properly) will reveal all repeated elements. If one starts assembling the most frequent l-mer (in order of decreasing multiplicities) in a genome into a de Bruijn graph, would all paths in the resulting graph represent distinct repeat families? The answer to the question is NO, since different subfamilies of the same repeat family may have slightly different versions of the same l-mer, with both variants appearing with high frequency. Aggregating this effect over different l-mer positions, there may be an exponentially large number of paths in the graph corresponding to a single repeat family.

   PROBLEM: Filter the list of frequent l-mers and assemble l-mers from the filtered list so that the resulting graph contains only one path for each repeat family.

   REQUIREMENT: Implementation skills.

   ASSIGNED PAPERS:

6. Finding highly diverged copies of known repeat families (improving repeat masking tools).

   The existing repeat masking tools (like RepeatMasker) may miss some highly diverged copies of repeats.

   PROBLEM: Complement the existing repeat masking tools by a program that finds such highly diverged copies.

   REQUIREMENT: Implementation skills.

   ASSIGNED PAPERS:

7. Finding unknown repeat families whose copies are highly diverged.

   Alu repeats are the most common type of repeats in human genome. 280 nucleotide long Alu repeats occur over one million times in human genome and comprise 10% of its entire length. Biologists believe that Alu repeats "invaded" the human genome 40-60 million years ago. Different copies of Alu repeats mutate and as a result, Alu repeats that we see today are approximately 15% different from the consensus sequence that they had 60 million years ago (therefore, two different Alu elements are approximately 15%+15%=30% different). With time passing by, the divergence between different copies of Alu repeats will be larger and larger.

   Since Alu repeats are relatively "young" it is not difficult to find them and to derive the original "master" Alu sequence. However, if Alu repeats were 100 million years old (and 25% diverged from the *unknown* consensus), the divergence between every pair of the Alu repeats would be high and it would be non-trivial to find out that they are all related or even present in the human genome. In other words, there will be 1 million "invisible" repeat copies in the human genome. One cannot rule out that there are indeed such "invisible" repeats in human genome. Could you find them?

   PROBLEM: Find some previously unknown (highly diverged) repeats families.

   REQUIREMENT: Algorithmic and implementation skills.

   ASSIGNED PAPERS:

8. Analysis of flanking regions of Alu repeat elements.

   Alu repeat elements are known to insert at TT/AAAA motifs (with mismatches allowed). Specifically, the genomic sequence TTAAAAXXXXX becomes TTAAAAXXXXXAluA_nAAAAXXXXX, where A_n is called the poly-A tail (typically 20-80bp long) and AAAAXXXXX is called the direct flanking repeat (typically 5-20bp long).

   PROBLEM.

   - Identify the direct flanking repeats of all Alu elements. Beyond the first four positions (which will be dominated by A), look for a nucleotide bias in each position. Construct "Alu insertion motif" (if it turns out to be strong enough).
   - Based on the Alu insertion paradigm described above, one would expect to see instances where a TT/AAAA insertion site is reused, leading to two Alu repeat elements (in the same orientation) separated only by a direct flanking repeat. How often does this happen? Is the number of instances of two closely spaced Alu repeat elements (in the same orientation) much larger than once would expect by chance if Alu elements inserted at random locations? Study the Alu subfamilies of the found "tandem Alus" and check whether they satisfy the proposed timeline of Alu subfamily classification (e.g., do the first Alu in the tandem repeats belong to a younger Alu subfamily than the second one?)
   - How many exact occurrences of TTAAAA are there in the human genome, and how many of these have ever experienced Alu insertions?

   REQUIREMENT: Implementation skills.

   REFERENCES: Batzer MA and Deininger PL, Nat. Rev. Genet. 2002, "Alu repeats and human genomic diversity".

   Dewannieux M et al., Nature Genetics 2003, "LINE-mediated retrotranposition of marked Alu sequences".

9. Analysis of trimeric LLR Alu elements.

   Most Alu elements are dimeric, represented by LR where L (left Alu monomer) and R (right Alu monomer) are highly homologous. It is known that there are a relatively small number of trimeric

LLR Alu elements in the human genome. One possible way they could have formed is via unequal homologous recombination between the right Alu monomer on one chromosome copy and the left Alu monomer (of the same Alu element) on the other chromosome copy, during meiosis. Letting $L = L_1L_2$ and $R = R_1R_2$, where the place recombination occurs is in between L_1 and L_2 (R_1 and R_2, resp.), this would lead to trimeric Alus of the form $L_1L_2R_1L_2R_1R_2$.

PROBLEM: How many instances of trimeric LLR Alu elements are there in the human genome? Are they predominantly of the form $L_1L_2R_1L_2R_1R_2$, verifying the recombination paradigm?

REFERENCES: Batzer MA and Deininger PL, Nat. Rev. Genet. 2002, "Alu repeats and human genomic diversity".

10. Comparison of repeats in human and chimpamzee.

    Recently sequenced chimpanzee genome opens a possibility to study recently introduced (or deleted) repeats and learn what factors promote repeat insertions/deletions. In particular, the following two step mechanism was recently proposed for Alu removal. First, another Alu is inserted close (20-100 nucleotides) to the existing Alu in the opposite orientation. The resulting pair of Alus has a tendency to form a duplex that is removed during the replication process. As a result, insertion of a new Alu in such case provokes a removal of both the new and old Alu elements.

    PROBLEM: Find all Alu elements that are present in human but absent in chimpanzee genome (and vice versa). Investigate the process of Alu removal/insertion and classify all Alus into (i) insertions in the human (chimpanzee) lineage and (ii) removals in the human (chimpanzee) lineage.

    REQUIREMENT: Implementation skills.

    REFERENCES: Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA. Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. Genome Res. 2001 Jan;11(1):12-27.

11. Quantification using LCMS

    Active proteins can be probed and quantified using liquid chromoatography based mass spectrometry. Digested peptides from a complex mixture of proteins are separated using liquid chromatography. As the peptides elute from the LC column, spectra are acquired. This collection of mass spectra (*LCMS map*) can be used to quantify active proteins.

    PROBLEM: Build tools to visualize and compare LCMS maps.

    REQUIREMENT: C++ implementation skills.

    RESOURCES: The OpenMS library should be helpful in building quick prototypes *http://open-ms.sourceforge.net/software.html#signal*.

12. Multiple alignment of RNA sequences

    RNA sequences are like DNA. However, they evolve under different constraints, which allow them to retain secondary structure even as the primary sequence diverges. This makes multiple alignment of RNA extremely challenging. The traditional approach is to model RNA as a stochastic context free grammar whose rules capture primary sequence as well as secondary structure constraints. In theory, this allows for the simultaneous deduction of multiple alignment, and structure. In practice, however, SCFGs have only had limited success. Here, we tackle a somewhat easier problem.

    PROBLEM: Given a collection of RNA sequences annotated with conserved structure, construct a multiple alignment that aligns the structural elements.

REQUIREMENTS: Algorithm and Implementation skills.

REFERENCES: Durbin, Eddy, Krogh, Mitchison (1998). Biological sequence analysis.

Bafna, Tang, and Zhang (2005) RNA consensus folding revisited. Preprint.

13. Aligning pseudo-knotted RNA to a sequence. The problem of aligning a sequence to an RNA structure is solved for the case when the structure has no pseudo-knots. Also, some recursive formulations are known that allow us to predict structure for simple pseudo-knots. The goal is to combine the two ideas.

    PROBLEM: Given an RNA sequence with known pseudo-knotted structure, align it to another sequence.

    REQUIREMENTS: Algorithmic, implementation skills.

    REFERENCES: T. Akutsu. Dynamic Programming algorithm for RNA secondary structure prediction with pseudoknots. Disc. Appl. Math, 104:45–62, 2000.

    Bafna, Muthukrishnan, and Ravi. Computing similarities between RNA strings. Combinatorial Pattern Matching Conference, 937:1–14, 1995.

14. Discovering orthologous conserved RNA sequences. Rivas and Eddy describe a tool, QRNA, for distinguishing conserved sequences as being regulatory, coding, or RNA. However, many conserved RNA sequences have diverged to the point where they may not show up as conserved. To start with, you must review the paper below, and determine why the authors miss many putative RNA. Also, the search for putative orthologs must be efficient, and filtering ideas might be used.

    PROBLEM: Given a pair of orthologous genomic regions (Ex: human and mouse), compute all pairs of sequences that encode ncRNA.

    REQUIREMENTS: Algorithmic, implementation skills.

    REFERENCES: (QRNA) Noncoding RNA gene detection using comparative sequence analysis. Elena Rivas and Sean R Eddy. BMC Bioinformatics, 2–8, 2001.

    (Filtering) Searching Genomes for non-coding RNA using FastR. Shaojie Zhang and Brian Haas and Eleazar Eskin and Vineet Bafna. (preprint).

15. *De novo* sequencing of paired spectra.

    Newer protocols in MS based quantification often result in pairs of MS2 spectra for the same peptide, but with differential post-translational modifications. The two spectra are very similar, with characteristic shifts due to the modifications. For example, if the modification was at the C-terminus of the peptide, all the prefix ions would remain unchanged, and match up in the two spectra, while all suffix ions would shift by a characteristic amount. Separating the prefix and suffix ions would make peptide sequencing trivial. However, this situation is complicated as the modification attaches to some 'unknown' position in the middle.

    PROBLEM: Given a pair of spectra corresponding to the modified and unmodified forms of the peptide, determine the peptide sequence.

    REQUIREMENT: Algorithmic/Implementation skills

    REFERENCES: Pevzner PA, Mulyukov Z, Dancik V, Tang CL.

    Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome Res. 2001 Feb;11(2):290-9.

    Bern and Goldberg (2005).EigenMS: de novo analysis of peptide tandem mass spectra by spectral graph partitioning. Preprint.