1 Logistics

The final 1 or 2 lectures of the class will be devoted to final presentation of the project. You are required to work on one of the projects described below. Checkpoints have been created to help you stay on track. They will not be graded separately, but you must meet the deadlines.

Checkpoints

C1: 11/3/05 Submit a 1-2 page written report with answers to the following questions:

1. Your project partner’s name, if any. Teams of 2 people are recommended, but 1 or 3 will also be accepted.
2. Your choice of the project. Choose from one of the following.
3. Answers to the C1 questions for the project.

C2: 11/10/05: Answers to the C2 part of the project.

C3: 11/24/05: Answers to the C3 part of your project. Schedule a meeting with the instructor to discuss final project.

C4: 11/29, 12/1: C4 part, and final presentation.

2 EST motifs

Goal: Identify motifs

Background: Prof. Macagno's lab works on neuronal development in the medicinal leech. Recently, they have generated a comprehensive collection of 3' ESTs from this organism. The data-set and an analysis of the genes will be published soon. Here, we ask the following question: Do the 3' UTR regions contain any interesting signals?

Requirements: Scripting language, Basic biology. This project needs some analysis, and it will help to have someone on your team with the requisite background.

Contact: Richard Rouse

Questions: Answer the following:

1. What is an EST? What is the Untranslated Region (C1)?
2. What is UTRdb. Write a short paragraph on the resources in this database (C1).
3. For the given data set, identify the coding boundaries of the 3' UTR in each EST, as follows (C2):
   (a) Blast the ESTs against a protein data-set.
   (b) Compute the distribution of the end-point of the alignments. Are the alignments high quality. Is the end-point tightly distributed?
   (c) Locate a STOP codon near the end-point of the alignment. Try all 3 frames.
(d) Compare your results to the annotations provided.

4. Create a new data-set of the untranslated region. Compute the distribution of sizes in it.

5. UTRdb is a database of motifs, and untranslated sequences found in the untranslated regions. Download UTRsite (the database of UTR motifs).

6. Write your own program to search the predicted untranslated regions with against UTRsite. Submit the result of your searching the first 20 ESTs (C3).

7. Search all of the data with UTRsite.

**Presentation Guidelines:**

1. Present slides on the methodology, with some illustrated examples.

2. Answer the following: What fraction of ESTs hit (significantly) only one reading frame? What fraction seemed to have a frame-shift?

3. You can ‘extrapolate’ the hits to get a putative end. Second, you can get an end by looking for a stop codon after the hit (in the same reading frame). Compute a distribution of the difference in the two end-points. Present 2 examples of outliers in this distribution, and give plausible reasons for the large difference.

4. Provide some statistics. How many UTR regions contain a motif? How many motifs are found in the leech data-set? For each motif, look up the function of the motif.

5. For each motif, cluster all of the ESTs that contain this motif. Many of them will be identical. For a non-redundant set of ESTs, are their protein sequence similar? What is the relation between the UTR motifs, and the protein function? You can use the output from the EST clustering team.

6. For the most "interesting" cases, find UTRs of homologous proteins from other species. Do they contain the same motif? You can use Genbank, or UTRGenome.

3 Searching genomes with mass spectra

**Goals:** the goal of this project is to construct a database of putative exons, to facilitate searching genomes with mass spectra.

**Contact:** Stephen Tanner

**Detailed description:**

1. What is an exon? What is an Intron? What is splicing? What is a splice site (C1)?

2. Search the web to get an example of alternative splicing (C1)?

3. Given a protein sequence, use Blast over the web to identify genomic regions that code for this protein sequence (C2).

4. Write your own splice site finder, or download one off the web. Given a genomic sequence, use the splice site finder to identify putative donor and acceptor sites. How many do you find in a 1Mb genomic region (C2)?

5. Write (C/C++) code for the following: given a genomic sequence, and a protein sequence, use blast local alignments, and splice site detection to compute the exon and intron coordinates exactly (C3).

6. Given a collection of protein sequences and a genome, identify exon-intron coordinates for all proteins. Construct a database of putative exon pairs. If the compute time is too large, run against chromosome 1 and 2.

7. Compile and run InsPecT, software for MS based identification, against the protein database

8. Report on the following (C4):
   
   (a) Number of protein sequences that were mapped to the genome.
(b) Number of locations (overlapping hits).
(c) The fraction of genome (chromosome) that is coding.
(d) Number of Exons mapped.
(e) Number of Branchpoints (exon boundaries that have alternative connections).
(f) For every gene identified, generate a protein sequence by translating the genomic sequence. Compute the % identity between the query protein, and the predicted protein. Plot a distribution of the % identity. Give plausible reasons why you do not get 100% identity.
(g) Number of peptides identified using MS.
(h) Number of peptides that cross splice-boundaries.
(i) Number of branch-points that are validated by peptides. In other words, how many cases do you have of peptides (identified using MS) that match 2 or more splice variants at a junction.