AM2: ISMB2006 Tutorial

Title: Genomes, Browsers and Databases: Tools for Automated Data Integration across Multiple Genomes

Topic Area:

- Database and Data Integration

Main Presenter:

- Title: Dr.
- Full name: Peter Schattner
- Affiliation: Department of Biomolecular Engineering, University of California, Santa Cruz
- Mailing Address: 820 Sea Spray Lane #301, Foster City, CA 94404, USA
- Email address: schattner@cse.ucsc.edu
- Telephone number: 1-650-574-7694
- Fax number: 1-650-574-7694
- Home page URL: http://www.spectelresearch.com/PSchattner/professional.html

- Teaching experience: Taught bioinformatics courses at University of California, Santa Cruz and California State University, Hayward
- Earlier tutorial presentations and feedback if any – give tutorial title, conference name, location, year:

"Perl and Bioperl: Tools for Automated Analysis of Biological Sequence Data". Tutorial presented at 2002 O'Reilly Bioinformatics Conference (Tucson, AZ), 2002 IEEE Bioinformatics Conference (Stanford, CA) and 2002 ISMB Conference (Calgary, Alberta). The O'Reilly tutorial sold out, with 120 registrants, six weeks prior to the conference. The ISMB tutorial received a 93% student approval rating (in contrast to a 78% rating for all tutorials at that ISMB).

50-word abstract:

The UCSC, Ensembl and NCBI genome databases integrate data from multiple, disparate sources in a uniform manner. However, developing automated queries to access these integrated databases has a considerable learning curve. Using realistic examples, participants will learn to design queries and programs enabling such automated analyses of genomic data.

Tutorial level:

I would prefer to rate the Tutorial level as "Intermediate", but if constrained to choose between Introductory and Advanced, I would classify the Tutorial as "Introductory".

Prior knowledge required:

The tutorial is self-contained; however, exposure to the UCSC, ENSEMBL or NCBI Genome Browsers would be helpful. Some experience with sequence similarity searching, multiple sequence alignment, and relational databases and programming in a language such as C or Perl would be useful for parts of the tutorial. No specific knowledge of biology is required; however, familiarity with the typical molecular biology sequence analysis tasks - such as retrieving sequences from databases, parsing database files and comparing EST or mRNA data to genomic sequences – will make it easier for the student to appreciate the advantages of the tools being presented.

Suitability of this tutorial for ISMB:
The number of molecular biology databases continues to explode. Presently, few problems in any area of genomic molecular biology can be addressed without analysing data stored in these databases. However, these databases are located in many different locations and often use non-standard data formats requiring specialized data parsers. As a result integrating and comparing data from multiple biological databases is difficult and tedious.

The genome databases at UCSC, Ensembl and NCBI offer solutions to this problem by integrating data from multiple databases in a uniform and standardized manner. However, effectively using these databases also has a considerable learning curve, especially if one wants to query multiple genomic regions in an automated manner rather than simply anaylzing individual genes via an interactive browser. This tutorial is intended to help students and researchers climb this learning curve more expeditiously.

The tutorial should be useful for both biology and bioinformatics students and researchers. Biologists will learn how to extract a wider range of relevant annotations for their genes of interest from the browsers. Bioinformaticians will learn how to access the underlying browser databases to perform automated, large scale queries across entire genomes. As important, both groups will gain an appreciation for the methods by which the browsers and their databases are constructed so that they are prepared to take advantage of new features and enhancements that are continually being incorporated into these important tools.

Profile of Presenter
My current research interests are focused on genome-wide identification and characterization of non-protein-coding RNA genes and cis-regulatory mRNA signals. This work has resulted in seven recent articles – published or under current review in major peer-reviewed journals – in which I am the lead and/or corresponding author.

My research has required extensive comparative genomic analysis using the UCSC and, to a lesser extent, the Ensembl, browsers and databases. In the process, I have worked closely with several of the developers of the UCSC Genome Browser. I also have been a key developer and worked closely with the other developers of the Bioperl project which is a principal software component underlying the Ensembl database.

Along with my research interests, I have focused my efforts on teaching and presenting bioinformatics concepts to a wide range of students and researchers. In addition to teaching regular university bioinformatics courses, I developed a tutorial presentation on Bioperl that was extremely well received at several conferences (see above for feedback information). I have also written the widely used Bioperl Tutorial as well as three other recent tutorial reviews on the biology and computational aspects of non-coding RNAs. One of these reviews is currently the #1 most downloaded article from the journal *Trends in Genetics:*
http://top25.sciencedirect.com/?journal_id=01689525

Tutorial Outline:

Note; The tutorial will discuss all three integrated genome browsers and, in particular, describe some of the tradeoffs between them. However, for pedagogic reasons, as well as because of time constraints, most of the presentation will focus on a single browser (specifically, the UCSC Browser).

Part I - Genome Browser Introduction (15 min)

What sorts of tasks are suited for the Genome Browsers?

A quick overview of the three genome browsers (UCSC, Ensembl, NCBI)

Part II –Browser Basics – using the UCSC Browser as an example (45 min)

Basic Track Types:

- Genes & gene predictions
- mRNAs and ESTs
- Expression and Regulation
- Comparative genomics
- Variations and Repetitive Sequences

Track Formats

Builds and Assemblies

Tracks and Builds: or Where did that track go?

An example: Finding SNPs, conserved subregions and expression patterns of the BRCA1 gene

Part III – More Advanced Browser Features (30 min)

BLAT vs Blast vs Blastz vs Multiz etc.

Blastz alignments – chains, nets, synteny and all that

Using “genome-test”– UCSC’s development browser

When to use other browsers (eg Ensembl or NCBI) or other tools

Part IV – Beyond the Browser, querying multiple genomic regions simultaneously (1 hr)

Browser File Formats –bed, psl, maf

The DAS approach for building browser tracks

Writing Custom Tracks

The Table Browser

The Galaxy Front end

An example: Finding promoters that are conserved and/or bind known transcription factors

Part V – Beyond the Browser, writing automated procedures for genome-wide queries (30 min)

Using the Browser Source Code Base

Local Browser Installation

Writing a C program to automatically extract database data

An example: Characterizing introns that host snoRNA genes

Finding more information about the browsers and their databases

In addition to the 3 hrs allocated above to the formal presentation, 30 minutes of the presentation time are allocated to discussion of the various topics presented.