

Brief description of the instructor(s) indicating the relevant qualifications and teaching experience

Peter Schattner, Ph.D., was trained as a physicist, but currently works as a computational biologist. He is a member of the computational-biology group at the University of California, Santa Cruz where his principal research interests are in computer-based methods for identifying non-protein-coding-RNA genes. He is also an active developer for the Bioperl project, having contributed several of its modules as well as having written the Bioperl tutorial. In his "spare time" he enjoys teaching the use of Perl and Bioperl in bioinformatics in venues ranging from the O'Reilly Bioinformatics Conference (http://conferences.oreillynet.com/cs/bio2002/view/e_sess/1948) to the Bioinformatics Program at the California State University, Hayward, CA. (<http://www.extension.csuhayward.edu/html/inf.htm>). His O'Reilly Bioperl tutorial sold out (with 120 registrants) six weeks prior to the conference.

Title

Perl and Bioperl: Tools for Automated Analysis of Biological Sequence Data

Expected goals, objectives and motivation of the tutorial

The rapid growth of sequence data in public and proprietary databases has created unprecedented opportunities to mine human and other genome databases for novel genes, proteins and biological pathways. In addition, with web-based user interfaces, these resources have become accessible to researchers with limited skills with computers. However, using these interactive interfaces becomes unwieldy when faced with complex comparative analyses involving hundreds or thousands of sequences and multiple databases.

Perl has emerged as the language of choice for the automated access and manipulation of bioinformatics data. However, while writing Perl programs is relatively easy, fully exploiting Perl's bioinformatics capabilities requires a level of programming experience and sophistication not common among biologists. To address this problem, the Bioperl Package - a set of object-oriented modules that implements common bioinformatics tasks - has been developed.

This tutorial describes Perl and Bioperl and their application to practical problems in molecular biology sequence analysis. The tutorial includes an overview of the principal features of Perl and Bioperl relevant to biology, followed by examples of how they can be applied to common bioinformatics tasks. Attention will also be paid to identifying bioinformatics problems for which Perl and Bioperl are not appropriate tools. By the end of the tutorial, participants will have a sense of what capabilities Perl and Bioperl can provide them in molecular biology research, as well as pointers to resources for acquiring the skills and knowledge they will need in order to take advantage of them.

Intended audience:

The tutorial is designed to be at an intermediate level. Experience with some computer language (eg C, java) is assumed. Some experience with Perl is helpful.

However, participants who are comfortable in another computer language should find the tutorial valuable even if they have little or no experience with Perl. Advanced knowledge of Perl - such as being able to write object oriented Perl - is not needed. No specific knowledge of biology is required to follow the tutorial. However, familiarity with the typical informatics tasks required by molecular biology sequence analysis - such as retrieving sequences from databases, converting sequence data files among the various standard formats, and parsing results of programs such as BLAST or HMMer - will make it easier for the student to appreciate the advantages of using the tools being presented.

Length: half or full day

Half day

Detailed outline of the presentation.

Part I - Overview of Perl and Bioperl and their Application in Bioinformatics

What bioinformatics tasks are suited to Perl?

What are the main capabilities of the Bioperl package?

Example 1 - Searching for promoter sequences

Example 2 - Aligning and "Blasting" protein sequences

When shouldn't you use Perl and Bioperl?

Part II - Selected Topics in Perl for Bioinformatics

Regular expressions and pattern matching

Obtaining perl objects - CPAN and Bioperl

Using objects in perl

Miscellaneous useful features of Perl

Debugging techniques in Perl

Part III - Getting started with Bioperl

Installing Bioperl

Overview to Bioperl's objects

Sequence objects: (Seq, PrimarySeq, LocatableSeq, LiveSeq, LargeSeq, SeqI)

Interface objects and implementation objects

Part IV - Using Bioperl

Accessing sequence data from local and remote databases

Transforming formats of database/ file records

Manipulating sequences

Searching for "similar" sequences - running and parsing Blast and HMMer

Running sequence alignment programs - Smith-Waterman, Clustalw, TCOffee

Manipulating / displaying alignments

Parsing results of programs that search for genes and other genomic DNA structures

Developing machine readable sequence annotations

Perl, Bioperl and XML

New and upcoming developments - Bioperl-db, Bioperl-gui, Structure objects

Part V - Learning more about Perl and Bioperl for Bioinformatics

Finding more information about specific Bioperl objects

Using sample code from the bptutorial script

Finding where a method is defined using the bptutorial script

Using the ptkdb debugger

Using the 't' and 'examples' directories

Recommended Books on Perl

Other Tools and Resources