Introduction to Bioinformatics

Bloomsbury Centre for Bioinformatics

PRACTICAL 1: BLAST and Sequence Alignment

Brief description of tutorial:

The EBI and NCBI websites, two of the most widely used life science web portals are introduced along with some of the principal databases: the NCBI Protein database, UniProt and the species specific databases organised under EnsemblGenomes. Both simple and more detailed examples of searching these databases using search terms and strings are considered. Analysis of the results and the refinement of search strings are discussed. The importance of utilising multiple databases to perform a thorough search is illustrated by the comparison of search results obtained from different databases.

BLAST searching is introduced and example BLASTN and BLASTP searches are performed. The importance of assessing the validity of the results obtained is explored with practical examples. More advanced BLAST searching is carried out and the selection of databases and the importance of various search parameters is discussed. Iterated BLAST searching using PSI-BLAST is introduced and the problems that might arise when analysing search results are highlighted.

The final exercise introduces sequence alignments, which are performed using CLUSTALW and T-COFFEE, and presents the issues that need to be considered to achieve optimal results.

Aims of session:

This practical aims to walk you through the process of text searching DNA and protein databases for sequence entries.

By the end of this practical you should:

- Be familiar with simple searches of DNA and protein databases for sequence entries.
- Common text query problems encountered when searching databases.
- Understand how to integrate sets of bioinformatic queries.
- Know about the main databases of biological information.
- Know how to query the databases by searching annotations or by using a query sequence.
- Have experience of using BLAST.
- Know how to extend the potential coverage of your searches using PSI-BLAST for iterated BLAST searches.
- Understand some of the potential problems you may encounter when using BLAST.
- Know how to perform and analyse a multiple sequence alignment.

Bloomsbury Centre for Bioinformatics

PRACTICAL 2: Protein Structure and Families

Brief description of tutorial:

The commonly used molecular visualisation tool RasMol is introduced and a short video tutorial demonstrates its key utilities. An exercise is carried out where a protein structure is examined in RasMol using some of these basic features, such as colouring by structure and selecting particular residues.

The PDB databank, the main repository for protein structural information, is briefly described and then searched with BLAST using an example protein sequence. The PDBSum website is studied and explored to demonstrate the information that can be retrieved.

Structural Classification is introduced and the major databases, CATH, SCOP and DALI, are described. Searches of CATH and SCOP are performed in various exercises and their similarities and differences are discussed. Some examples of proteins with different structures are provided for examination using PDBSum, CATH and SCOP to indicate the diversity of protein structure.

Secondary structure prediction is introduced and the widely used PSIPRED server is utilised to examine a mystery protein sequence. The DISOPRED server is used to estimate the level of disorder in the mystery sequence. The fold recognition method pGenThreader is also applied in the study of the mystery protein. Secondary structure prediction of membrane proteins is discussed. The corresponding specialised methods, TMHMM, which predicts transmembrane (TM) helices and MEMSAT, which performs both TM helix prediction and topological prediction, are illustrated using an example.

Finally to bring together the various tools and methods presented in the practical session, analysis of a couple of further mystery protein sequences is performed.

Aims of session:

This practical teaches you about what resources and utilities are available for predicting the structure of protein sequences, finding known structures of protein sequences and examining structural classifications of proteins.

By the end of this practical you should:

- Be familiar with using RasMol for viewing protein structures.
- Know how to go about finding structural information for protein sequences.
- Be familiar with the principal two structural classification resources: SCOP & CATH.
- Know about the difficulties of classifying structures and the differences between classifications assigned.
- Know about the options for predicting protein secondary structure.
- Understand the limitations of current approaches.

Bloomsbury Centre for Bioinformatics

PRACTICAL 3: Homology Modelling

Brief description of tutorial:

Homology Modelling is introduced and the main steps performed to produce a homology model are outlined and the methodology is discussed. The PDB (introduced in Practical 2) is looked at in more detail via a short online video tutorial. The Swiss-Model server is introduced and Swiss-Model is used in the Automated Mode to produce a homology model of a given protein. Interpretation of the results obtained from the server is discussed.

The commonly used molecular visualisation tool Swiss-PDB Viewer is introduced by two short video tutorials. It is then utilised to check the quality of the homology model produced earlier in the practical session using the Swiss-Model server.

The popular molecular visualisation tool PyMol is briefly demonstrated and exercises enable hands-on examination of a protein structure using tools available in the software. The use of PyMol for producing publication quality images is introduced.

Finally manual homology modelling using the script based software Modeller is briefly introduced.

Aims of session:

This practical is about predicting 3-dimensional protein structure using the best method currently available: homology modelling. In this procedure, the detailed 3D structure of an uncharacterised query sequence is predicted using the structure of a related protein with a similar sequence.

By the end of this practical, you should:

- Understand the basis of homology modelling.
- Have knowledge of the important factors for making a good prediction.
- Know how to use a popular resource for building models, SWISS-MODEL.
- Have experience interpreting the output of SWISS-MODEL.
- Know about some basic model quality checking tools.
- Be familiar with the Swiss-PDB Viewer (SPDBV) package for modelling.
- Have experience in using PyMol for the visualisation of protein structures and for producing high quality images of proteins.
- Be aware of script based manual homology modelling using software such as Modeller.

Bloomsbury Centre for Bioinformatics

PRACTICAL 4: Microarray Analysis

Brief description of tutorial:

Microarray technology is introduced with a description of Affymetrix GeneChip Arrays. Microarray data is obtained from public repositories. The Microarray analysis workflow is outlined. A significant portion of the tutorial is based on the use of the popular Bioconductor package, which requires some knowledge of R commands, for the analysis of microarray data. Bioconductor and R are introduced and the microarray analysis workflow is illustrated using a working example of precomputed data to demonstrate basic commands and applications. Quality Control (QC) analysis, Normalisation, Differential Gene Expression Analysis and Statistical Analysis of Differential Gene Expression are discussed.

Gene annotation is described and annotation information retrieval using NetAffx and the Gene Ontology website is introduced. Further analysis of gene function is demonstrated with the use of the webtools DAVID and Genmapp. The application of microarray analysis to disease classification is demonstrated with the web-based clustering tool EpClust.

Aims of session:

This practical provides a brief exploration of the process of microarray data analysis. Starting with the raw intensity values obtained from scanned arrays, it covers:

- Simple Quality Control (QC) steps for Affymetrix data.
- Data normalisation.
- Statistical Analysis of differential gene expression.
- Obtaining Gene Ontology (GO) definitions.
- Introduction to freely available tools for functional annotation and analysis, text-mining and clustering.
- Application of microarrays and clustering methods to disease classification.

Bloomsbury Centre for Bioinformatics

PRACTICAL 5: Genomics and Phylogenetics

Brief description of tutorial:

The Ensembl Genome Browser is introduced and methods to search, view and interpret genomic data and the search results obtained are shown using exercises. The GENSCAN webserver is used for gene prediction and the results are then compared with those obtained using Ensembl.

Phylogenetic analysis is introduced. Phylogenetic trees are viewed and interpreted using the Ensembl Gene Tree utility. The construction of phylogenetic trees is described. Other phylogenetic methods such as maximum parsimony, maximum likelihood and bootstrapping are introduced.

Data mining using the Ensembl BioMart tool is performed to illustrate how microarray data available in the Ensembl database can be searched and analysed. Genes with a disease association are used as examples.

Aims of session:

This aim of this session is to introduce you to the bioinformatic tools and resources available on the web for working with genomic data.

By the end of the practical you should:

- Be familiar with the Ensembl genome browser system.
- Know how to query genomic data stored in Ensembl using text and sequence searches, and how to interpret the results.
- Understand the basic concepts and terminology of phylogenetics.
- Know how phylogenetic trees are built and the problems you can encounter during phylogenetic analysis.
- Be familiar with using BioMart for mining the Ensembl database.