Bioinformatics for biomedicine

Summary and conclusions. Further analysis of a favorite gene

Lecture 8, 2006-11-07

Per Kraulis

http://biomedicum.ut.ee/~kraulis

Themes in bioinformatics

1. Databases
2. Sequences
3. Sequence search
4. Sequence, evolution, function
5. Protein 3D structure
6. Sequence alignment
7. Annotation
8. Gene expression; data analysis
9. Pathways and processes

1. Databases

- Data models
  - Domain
    - Included, excluded
  - Central data object(s)
  - Relations
- Database policy
  - Manually curated vs. automated
  - Updates
  - Access, licenses, copyright

2. Sequences

- Sequence databases
  - Nucleotide, protein
  - Annotation
  - Cross-references, links
- Sequence analysis
  - Features
  - Similarities
  - Phylogenetics

3. Sequence search

- Sequence search
  - BLAST, FASTA
  - Smith-Waterman
- Sequence patterns
  - Regular expressions
    - Prosite
  - Hidden Markov Models (HMMs)
    - Pfam
  - PSI-BLAST, PHI-BLAST, HMMER

4. Sequence, evolution, function

- Sequence and evolution
  - Sequence similarity
  - Homology
- Sequence and function
  - Domains
  - Activity, enzyme, binding
- Function and evolution
  - Orthologs: Speciation event
    - Similar function, presumably
  - Paralogs: Duplication event
    - Divergent function, presumably

5. Protein 3D structure

- Protein sequence and 3D structure
  - Sequence determines structure
- Structure and function
  - Strongly conserved
- Structure prediction
  - Folding problem
  - Modelling: using similarity
- Structural features
  - Folds and domains

6. Sequence alignment

- Sequence alignment
  - Part of sequence search
  - Required for 3D model from template
  - Quality depends on similarity
- Multiple sequence alignment
  - Heuristic algorithms required
  - Hard to obtain optimal solution
  - Phylogenetics

7. Annotation

- Annotation: properties, features,…
- Association by guilt
  - Sequence similarity
  - Behavioral similarity
    - Gene expression
    - Proteomics
    - Binding, physical association
- Gene Ontology
  - Controlled vocabulary of keywords

8. Gene expression; data analysis

- Gene expression
  - EST, SAGE, microarrays
  - Experimental design
    - Time course
- Data analysis
  - Normalization
  - Clustering
  - Statistics
  - Visualization

9. Pathways and processes

- Gene activity
  - Protein activity and interactions
  - Expression as proxy
- Pathways
  - Metabolism
  - Signaling and regulation
- Biological processes
  - Temporal and spatial
  - Hierarchy: different levels and scales

Bioinformatics: The future

- More complete genomes
  - Phylogenetics
- Functional genomics
  - Annotation, experimental design, integration
- Pathways
  - Current DBs incomplete
  - Data model?
- Processes
  - How to model?
  - System biology; towards prediction

Bioinformatics on the web 1

- EBI www.ebi.ac.uk
  - Site to be modified 11 Dec 2006!
  - Databases
    - EMBL: Nucleotide sequences
    - UniProt: Protein sequences, annotation, literature
    - IntAct: Protein interactions
    - ArrayExpress: expression data
  - Tools
  - 2Can: Bioinformatics educational resource
  - Research groups

Bioinformatics on the web 2

- NCBI www.ncbi.nlm.nih.gov
  - Databases
    - GenBank, RefSeq
    - Proteins
    - OMIM, Taxonomy
    - PubMed
  - Bookshelf: Biology textbooks on-line
  - Tools
    - BLAST, Entrez

Bioinformatics on the web 3

- Ensembl www.ensembl.org
  - Eukaryotic genomes
    - Nucleotide sequence, genes, transcripts, proteins
  - Databases and tools
- Vega vega.sanger.ac.uk
  - Curated eukaryotic genomes
- ExPASy www.expasy.org
  - UniProt (Swiss-Prot & TrEMBL)
  - Databases and tools

Bioinformatics on the web 4

- GeneCards www.genecards.org
  - Human genes
  - Integrated database: Other DBs used
- GeneLynx www.genelynx.org
  - Human, rat, mouse
  - Links for genes to other DBs
- Google
  - Now several useful DBs indexed!
  - Google Scholar http://scholar.google.com/

Bioinformatics on the web 5

- SGD Saccharomyces genome DB www.yeastgenome.org
- BDG Drosophila genome DB www.fruitfly.org
- FlyBase Drosophila genome DB flybase.bio.indiana.edu
- MGI Jackson lab mouse genome DB www.informatics.jax.org

Bioinformatics on the web 6

- PDB 3D biomolecular structures www.rcsb.org/pdb
- 3D structural motifs hierarchy http://scop.mrc-lmb.cam.ac.uk/scop/
  - Manual curation
- 3D structure classification www.cathdb.info
  - Automated curation

Bioinformatics on the web 7

- KEGG www.genome.jp/kegg
  - Pathways, metabolic and signaling
  - Started with human and eukaryotes
- BioCyc www.biocyc.org
  - Pathways, metabolic
  - Started with prokaryotes
- Reactome www.reactome.org
  - Pathways, signaling, reactions
  - Started with human

Bioinformatics on the web 8

- Biological processes
  - Several dedicated to specific processes
  - Educational in nature
  - No developed data models
- Systems biology
  - www.systemsbiology.org (Seattle)
  - www.biochemweb.org/systems.shtml

Further reading 1

- Bioinformatics. Genes, Proteins & Computers
  - C.A. Orengo, D.T. Jones & J.M. Thornton
  - 320 pp
  - BIOS Scientific Publishers Limited, 2003
  - ISBN 1-85996-054-5
- Bioinformatics. Sequence and Genome Analysis
  - D.W. Mount
  - 692 pp
  - Cold Spring Harbor Lab Press, 2004
  - ISBN 0-87969-712-1

Further reading 2

- Sequence – Evolution - Function
  - E.V. Koonin & M.Y. Galperin
  - 488 pp
  - Springer, 2002
  - ISBN 1-4020-7274-0
  - NCBI Bookshelf http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call