Bioinformatics for biomedicine

Seminar: Sequence analysis of a favourite gene

Lecture 5, 2006-10-17

Per Kraulis

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

Course design

1. What is bioinformatics? Basic databases and tools
2. Sequence searches: BLAST, FASTA
3. Multiple alignments, phylogenetic trees
4. Protein domains and 3D structure
5. **Seminar: Sequence analysis of a favourite gene**
6. Gene expression data, methods of analysis
7. Gene and protein annotation, Gene Ontology, pathways
8. Seminar: Further analysis of a favourite gene

From the previous lecture

- 3D structure
  - Relation to function
  - Methods of determination
  - Databases
- Domains
  - Definition
  - Relation to function
  - Databases

3D structure and sequence, 1

- Sequence determines structure
- Anfinsen’s experiments: folding
- 3D structure prediction should work!
  - First principles (physics)
    - Extremely difficult
    - Only special cases
  - By similarity
    - Works reasonably
    - Depends on many factors

3D structure and sequence, 2

- Similar sequence -> similar structure
  - General statement
  - No “real” counterexamples
    - But a designed, extreme case exists
- A single 3D structure represents a family
  - Depending on sequence similarity
- Many sequences “allowed” for a structure

3D structure and sequence, 3

- Many sequences “allowed” for a structure
  - Sequence divergence possible over time

- Structure more conserved that sequence
  - Sequence may diverge beyond recognition
  - Structure may still be similar
  - Apparently unrelated sequences may form similar structures!

3D structure and sequence, 5

- Protein design
  - Target: A given structure
  - Specify a sequence that folds into it
- Some success in recent years
  - First principles approach
    - Computational prediction
  - Evolutionary approach
    - In vitro evolution experiments

3D structure and sequence, 4

- Example: TIM barrel
  - TIM = Triose isomerase
  - Example in PDB: 1TRI
  - Symmetrical, 8 alpha-beta units
  - Enzymes
    - Wide range of reactions, substrates
- Some TIM barrels have no significant sequence similarity
  - Sequence divergence?
  - Structure convergence?

Structure database

- PDB database
  - At RCSB: http://www.rcsb.org/pdb/
  - Via UniProt: http://www.ebi.uniprot.org/
- Many different 3D viewers
  - KiNG (uses Java)
  - RasMol (requires installation)

Unknown sequence?

- Given unknown sequence, what to do?
  - Set up a check list
    - Never optimal

- Note: Unrealistic exercise
  - "Premier tool of analytic chemistry: the phone"
  - Consider known facts before analysis
  - Consider the underlying problem

Check list, outline

- Where to search first?
  - UniProt
  - Pfam
  - Ensembl
- Function?
  - Annotation
  - Similarities
  - Domains, structure

Unk1

- File “unk1-seq.txt”
- Human protein
- No other info

Unk1 = Caspase 8

- Protease
  - Cysteine peptidase C14
  - Cascade, activation of others
- Hormone-triggered
  - TNF (tumor necrosis factor)
- Multicellular organisms
  - Apoptosis (programmed cell death)
  - Development, cancer

Unk2

- File “unk2-seq.txt”
- From C. elegans
  - Result of genetic screen
- Orthologue in human?

Unk2 = nhr-25

- Nuclear receptor (NR) family
  - Transcription factor
  - DNA-binding domain
  - Hormone-receptor domain
- Orphan NR
  - Unknown ligand; maybe none
- Multicellular organisms
  - Developmental processes
    - Moulting in insects

Unk3

- File “unk3-seq.txt”
- Mouse protein
- No other info

Unk3 = GPR141

- GPCR
  - G-protein coupled receptor
  - 7TM, seven transmembrane helices
- Orphan GPCR
  - Unknown ligand
- Human ortholog, Ensembl search
  - http://www.ensembl.org/Multi/blastview

Task for this week

- Perform analysis of sequences
  - http://biomedicum.ut.ee/~kraulis/bioinfo_bi
  - unk4-seq.txt (Danio rerio, zebrafish)
  - unk5-seq.txt (human, fragment)
- Organism?
- Protein family?
- Structure and function?
- Human medical relevance?