Bioinformatics for biomedicine

Protein domains and 3D structure

Lecture 4, 2006-10-10

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

Multiple alignment

- Example: ras proteins
- FASTA input file available at course site
- http://www.ebi.ac.uk/Tools/sequence.html
  - ClustalW, MUSCLE
- http://msa.cgb.ki.se/cgi-bin/msa.cgi
  - Kalign

Proteins and drug action

- Most drugs act via proteins
  - Binds to the protein
  - Inhibits or activates it

- “Drug targets”
  - Small subset of all proteins
  - Certain protein families
  - 3D structural basis

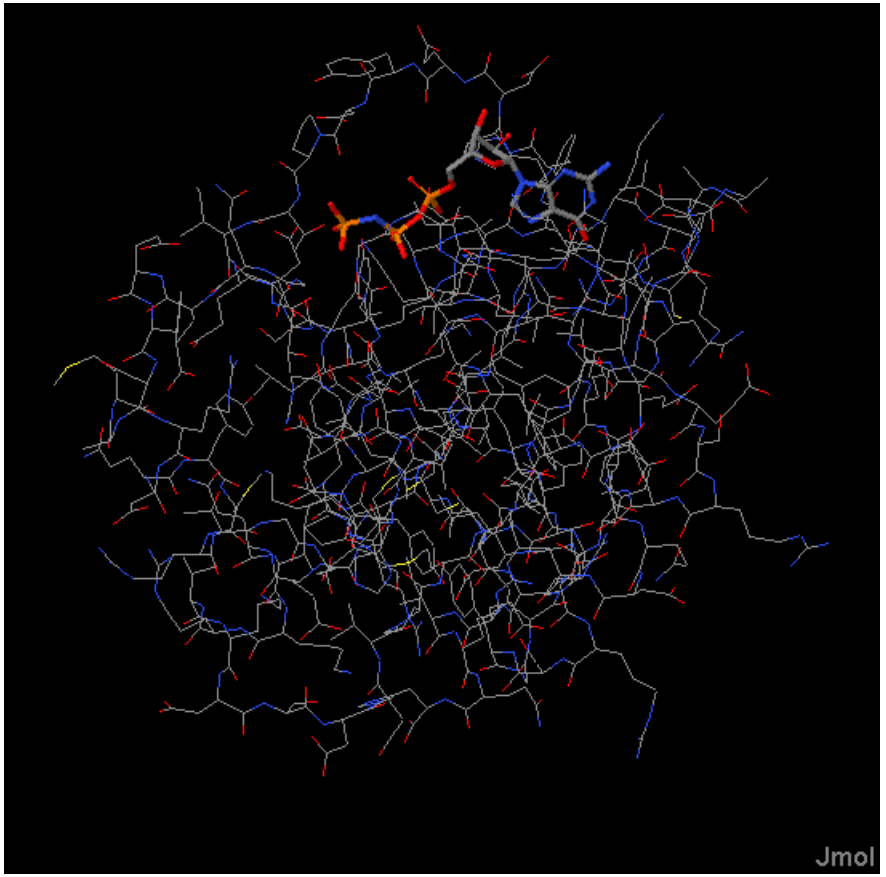
3D structure and protein function

- ‘Native’ state
  - Functional state
  - Well-defined 3D structure (usually)
  - Folded state

- Denaturing a protein inactivates it
  - Heat, salt, solvents… (boiling an egg)
  - Undefined 3D structure; a mess
  - Unfolded state

3D structures are complicated

- 1000's of atoms
- Hard to see anything
- But: details essential for understanding



Levels of 3D structure

- Primary
  - Protein sequence
- Secondary
  - Alfa helix, beta sheet; hydrogen bonds
- Tertiary
  - Overall 3D structure; “fold”
- Quaternary
  - Complex between biomolecules

Simplified view of 3D structure

- Schematic view
- Peptide chain
- Secondary structure
  - Alfa helix
  - Beta strands
- Ligands
  - Metal ions
  - Cofactors
  - Drugs

Structure database

- PDB database
  - 3D coordinates for structures
    - Protein, DNA, RNA, complexes
  - Experimentally determined structures
- At RCSB: http://www.rcsb.org/pdb/
- Via UniProt: http://www.ebi.uniprot.org/

- Many different 3D viewers…

Experimental determination

- X-ray crystallography
  - Atomic resolution
  - Small and large molecules (ribosome)
- NMR
  - Small, soluble proteins
- Electron microscopy
  - Not atomic resolution
  - Very large molecules/particles

X-ray crystallography 1

- Requires protein crystal
  - Size 0.1-1.0 mm
  - Possibly with ligands, other proteins
  - Optimize conditions (salt, solvents,…)

- Recent technological advances
- Structural Genomics Initiatives

X-ray crystallography 2

- Quality?
- Resolution
  - How much data was available?
  - Given in Ångström
    - Good: 2.0 Å or less
    - Bad: 3 Å or higher
- Refinement
  - How well do atomic coordinates agree with data?
  - R-value
    - Should be 20% or less

X-ray crystallography 3

- Is crystal state relevant?
- Short answer: Yes
  - High solvent content; similar to cell plasma
  - Comparisons indicate OK
- But: occasional relevant differences
  - Surface differences; loops
  - Details of ligand binding

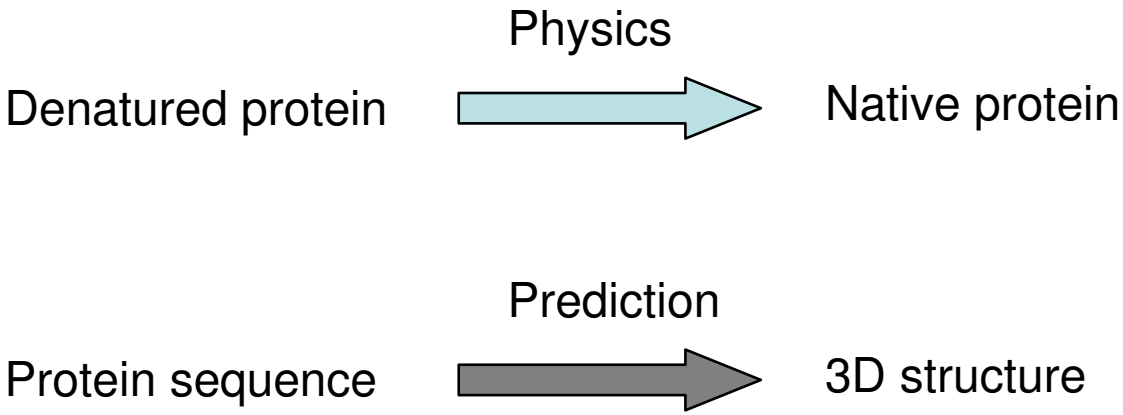
Protein structure and prediction

- Secondary structure prediction
  - Typically 70-80% accurate
  - Not as good as it sounds
  - Always use experimentally derived data, when available
- 3D structure modelling
  - Based on known structure: predict similar
  - OK when sequence >90 % similar
  - Depends, when 50-90% similar
  - Problematic, when <50% similar

The folding problem 1

- 3D structure encoded in protein sequence
  - Anfinsen 1960s
  - Small proteins refold spontaneously
  - Synthetic peptides fold
    - Activity as for ‘natural’ protein
- Assisted folding does not change this
  - Chaperones: catalysts for folding
  - Needed in cell environment (thick soup)

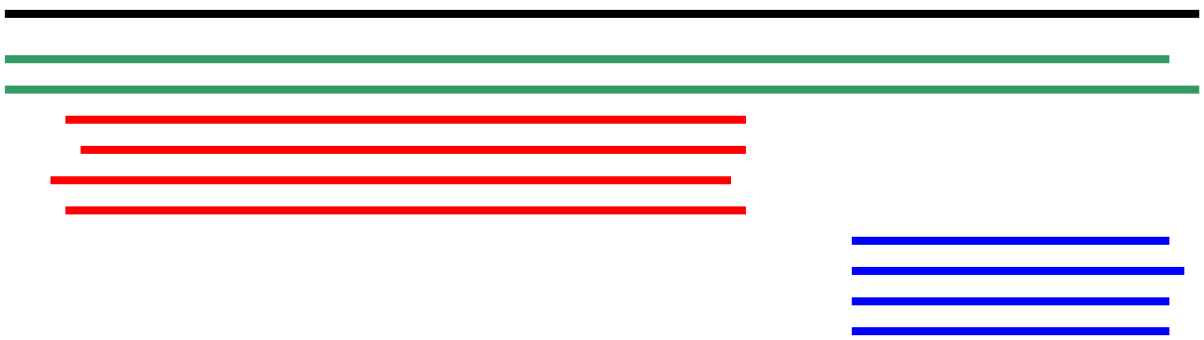
The folding problem 2



- Protein sequence determines 3D structure
- 3D structure should be predictable
- But: notoriously hard (unsolved) problem

Protein domains 1

- Proteins appear to be modular
- 3D structure: pearls on a string
- Sequence: partial sequence similarity



Protein domains 2

- Proteins are modular
  - Particularly in eukaryotes
- A part of sequence as a unit
- The units are called domains
  - Partial gene duplication

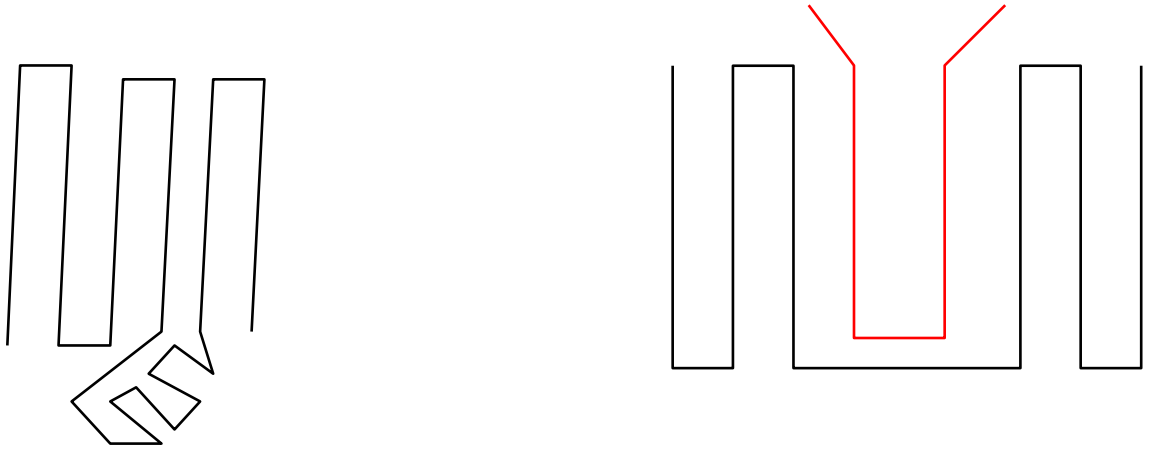
- Domain families
  - Proteins containing a particular domain

Protein domains 3

- Sequence-based domain definitions
- Pfam
  http://www.sanger.ac.uk/Software/Pfam/
  - Based on Hidden Markov Models (HMMs)
    - Statistical method to detect patterns
    - Sensitive
- Other domain databases
  - Structural: CATH http://www.cathdb.info/

Protein domains 4

- Domains from structure or sequence?
  - Usually very similar results
  - But some differences
    - Sequence region inserted
    - Structure formed from different parts of sequence



Protein families in medicine

- “Druggable” proteins
- Contains domains known to bind drugs
  - Small subset of proteins
- Usually: binds small molecules naturally
  - Enzymes
  - Receptors
- Protein-protein interactions very difficult
  - Much work, few results

Example: Nuclear receptors

- Transcription factors
  - 3D structures available for several
- Small-molecule ligands
  - Estrogen, progesteron,…
- Androgen receptor
  - ANDR_HUMAN, P10275
- Several drugs
  - Tamoxifen

Example: 7TM receptors

- 7TM = Seven transmembrane helices
- Receptors
  - CNS, other systems
- Maybe 50% of drugs act on these
- Integral membrane protein
  - Very hard for X-ray crystallography
  - No good 3D structure!

7TM: rhodopsin

- Light detector protein in eye
- 3D structure determined!
- Serves as model for other 7TM proteins
  - Known to be very approximate
- UniProt: OPSD_BOVIN, P02699
  - Retinal ligand shows binding pocket
  - In some 7TM proteins: different binding site

Example: protein kinases

- Regulatory proteins, signalling
  - Cancer
- Much recent work
  - Few drugs on market, yet
- Drug design strategy
  - Compete with ATP in its pocket
  - Specificity?
- UniProt: ABL1_HUMAN,P00519
  - PDB: 2F4J