Bioinformatics master course

DNA/Protein structure-function analysis and prediction

Lecture 1: Protein Structure Basics (1)

Centre for Integrative Bioinformatics VU (IBIVU)
Faculty of Exact Sciences / Faculty of Earth and Life Sciences
DNA/Protein structure-function analysis and prediction

**SCHEDULE**

http://www.few.vu.nl/onderwijs/roosters/rooster-vak-januari07.html

http://www.few.vu.nl/onderwijs/roosters/rooster-vak-voorjaar07.html

Centre for Integrative Bioinformatics VU (IBIVU)
Faculty of Exact Sciences / Faculty of Earth and Life Sciences
The first protein structure in 1960:
Myoglobin (Sir John Kendrew)
Protein Data Bank

Primary repository of protein tertiary structures

http://www.rcsb.org/pdb/home/home.do
Dickerson’s formula: equivalent to Moore’s law

\[ n = e^{0.19(y-1960)} \]

where \( y \) is the year.

On 27 March 2001 there were 12,123 3D protein structures in the PDB: Dickerson’s formula predicts 12,066 (within 0.5%)!
Protein primary structure

20 amino acid types

A generic residue

Peptide bond

SARS Protein From Staphylococcus Aureus

1 MKYNNHDKIR DFIIEAYMF RFKKVKPEV
31 DMTIKEFILL TYLFHQQENT LPFKKIVSDL
61 CYKQSDLVQH IKVLVKHSYI SKVRSKIDER
91 NTYISISEEQ REKIAERTVL FDQIIKQFNPL
121 ADQSESQMIP KDSKEFLNLML MYTMVFKNII
151 KKHHTLSFVE FTILAITTSQ NKNIVLLKDL
181 IETIHKYPQ TVRALNHLKK QGYLKERST
211 EDERKILIHMD DDAQODHAEOQ LLAQVNQLLA
241 DKDHLHLVFE
Protein secondary structure

Alpha-helix

Beta strands/sheet

SARS Protein From Staphylococcus Aureus

1 MKYNNHDKIR DFIIEEAYMF RFKKVKPEV DMTIKEFILL TYLFHQQENT
   SHHH HHHHHHHHH HHHHHHTTT SS HHHHHHH HHHHS S SE
51 LPFKKIVSDL CYKQSDLVQH IKVLTGHYYT SKVRSSKIDER NTYISISEEQ
   EEHHHHHHHS SS GGTHHH HHHHHTTT EEEE SSSTT EEEE HH
101 REKIAERVTL FDQIIKQFNL AQSESQMIK KDSKEFLNIM MYTMYFKNII
   HHHHHHHHH HHHHHHHHH HHHHSSS S HTT S S SHHHHHHH HH
151 KKLTLTSLFVE FTILAIITSQ KNIVLKLKDL IETIHKPQ TVRALNNLKK
   HHH SS HH HHHHHHHHTT TT EHHHH HHHSSS HH HHHHHHHHH
201 QGYLIERST EDERKILIHM DDAQDHAEOQ LLAQVNLALLA DKDLHLIYVFE
   HTHHSEE E S SSTT EEEE HHHHHHHHH HHHHHHHTS SS TT SS
Protein structure hierarchical levels

PRIMARY STRUCTURE (amino acid sequence)
VHLTPEEKSAVTALWGKVNVD
EVGGGEALGRLLVYPWTQRFF
ESFGDLSTPDAMGNPKVKAH
GKKVLGAFSDLHLDNLKGT
ATLSELHCDKLHDPENFRLLG
NVLVCVLAHFGKEFTPVQAA
YQKVVAGVANALAHKYH

SECONDARY STRUCTURE (helices, strands)

TERTIARY STRUCTURE (fold)

QUATERNARY STRUCTURE (oligomers)
Protein folding problem

Each protein sequence “knows” how to fold into its tertiary structure. We still do not understand how and why

**PRIMARY STRUCTURE (amino acid sequence)**

VHLTPEEKSAVTALWGKVNVD
EVGGEALGRLLVVYPWTQRF
ESFGDLSTPDAMGNPCKAH
GKKVLGAFSDGLAHLDNSLKGT
ATLSELHCDKHLVDPENFRLLG
NVLVCVLAHGFKEFTPPVQAA
YQKVVGVAHALAKYH

**SECONDARY STRUCTURE (helices, strands)**

1-step process

2-step process

The 1-step process is based on a hydrophobic collapse; the 2-step process, more common in forming larger proteins, is called the framework model of folding
Fig. 2.13. Examples of folds in globular proteins: (a) globin fold in α protein myoglobin (PDB:1MBN), (b) β sandwich in β protein immunoglobulin (PDB:7FAB), (c) TIM barrel in α/β protein triose phosphate isomerase (PDB:1TIM), and (d) a fold in α+β protein ribonuclease A (PDB:7RSA).
Globin fold
α protein
myoglobin
PDB: 1MBN
β sandwich
β protein
immunoglobulin
PDB: 7FAB
TIM barrel
α / β protein
Triose phosphate
IsoMerase
PDB: 1TIM
A fold in α + β protein ribonuclease A
PDB: 7RSA

The red balls represent waters that are ‘bound’ to the protein based on polar contacts
An α helix has the following features:
- every 3.6 residues make one turn,
- the distance between two turns is 0.54 nm,
- the C=O (or N-H) of one turn is hydrogen bonded to N-H (or C=O) of the neighboring turn -- the H-bonded N atom is 4 residues up in the chain.

(a) ideal right-handed α helix. C: green; O: red; N: blue; H: not shown; hydrogen bond: dashed line. (b) The right-handed α helix without showing atoms. (c) the left-handed α helix (rarely observed).
A β sheet consists of two or more hydrogen bonded β strands. The two neighboring β strands may be **parallel** if they are aligned in the same direction from one terminus (N or C) to the other, or **anti-parallel** if they are aligned in the opposite direction.
Crystal Structures of Some Outer Membrane Proteins

FhuA  (Ferguson et al., 1998)
FepA  (Buchanan et al., 1999)
OmpA  (Pautsch & Schulz, 1998)
But remember there are homologous relationships at very low identity levels (<10%)!
RMSD of backbone atoms (Å) vs. % identical residues in protein core

Chotia & Lesk, 1986
RMSD: Two superposed protein structures
(with two well-superposed helices)

Root mean square deviation (RMSD) is typically calculated between equivalent Ca atoms

Red: well superposed
Blue: low match quality

C5 anaphylatoxin -- human (PDB code 1kjs) and pig (1c5a)) proteins are superposed
Burried and Edge strands

Parallel β-sheet

Anti-parallel β-sheet
Secondary structure hydrophobicity patterns

**ALPHA-HELIX:** Hydrophobic-hydrophilic 2-2 residue periodicity patterns

**BETA-STRAND:** Edge strands, hydrophobic-hydrophilic 1-1 residue periodicity patterns; buried strands often have consecutive hydrophobic residues

**OTHER:** Loop regions contain a high proportion of small polar residues like alanine, glycine, serine and threonine.

*The abundance of glycine is due to its flexibility and proline for entropic reasons relating to the observed rigidity in its kinking the main-chain.*

*As proline residues kink the main-chain in an incompatible way for helices and strands, they are normally not observed in these two structures (breakers), although they can occur in the N-terminal two positions of α-helices.*
Flavodoxin fold

5(βα) fold
Flavodoxin family - TOPS diagrams
(Flores et al., 1994)

To date, all α/β structures deposited in the PDB start with a β-strand!
Protein structure evolution

Insertion/deletion of secondary structural elements can ‘easily’ be done at loop sites
Protein structure evolution

Insertion/deletion of structural domains can ‘easily’ be done at loop sites.
A domain is a:

- Compact, semi-independent unit (Richardson, 1981).
- Stable unit of a protein structure that can fold autonomously (Wetlaufer, 1973).
- Recurring functional and evolutionary module (Bork, 1992).

“Nature is a tinkerer and not an inventor” (Jacob, 1977).
Identification of domains is essential for:

• High resolution structures (e.g. Pfuhl & Pastore, 1995).
• Sequence analysis (Russell & Ponting, 1998)
• Multiple alignment methods
• Sequence database searches
• Prediction algorithms
• Fold recognition
• Structural/functional genomics
Domain connectivity

Figure 1: Scheme illustrating different types of connectivity in multidomain structures (left) and their sequences (right).
Domain size

• The size of individual structural domains varies widely from 36 residues in E-selectin to 692 residues in lipoxygenase-1 (Jones et al., 1998), the majority (90%) having less than 200 residues (Siddiqui and Barton, 1995) with an average of about 100 residues (Islam et al., 1995).

• Small domains (less than 40 residues) are often stabilised by metal ions or disulphide bonds.

• Large domains (greater than 300 residues) are likely to consist of multiple hydrophobic cores (Garel, 1992).
Domain characteristics

•Domains are genetically mobile units, and multidomain families are found in all three kingdoms (Archaea, Bacteria and Eukarya).

•The majority of proteins, 75% in unicellular organisms and >80% in metazoa, are multidomain proteins created as a result of gene duplication events (Apic et al., 2001).

•Domains in multidomain structures are likely to have once existed as independent proteins, and many domains in eukaryotic multidomain proteins can be found as independent proteins in prokaryotes (Davidson et al., 1993).
Domain fusion

Genetic mechanisms influencing the layout of multidomain proteins include gross rearrangements such as inversions, translocations, deletions and duplications, homologous recombination, and slippage of DNA polymerase during replication (Bork et al., 1992).

Although genetically conceivable, the transition from two single domain proteins to a multidomain protein requires that both domains fold correctly and that they accomplish to bury a fraction of the previously solvent-exposed surface area in a newly generated inter-domain surface.
Domain fusion example

Vertebrates have a multi-enzyme protein (GARs-AIRs-GARt) comprising the enzymes GAR synthetase (GARs), AIR synthetase (AIRs), and GAR transformylase (GARt) 1.

In insects, the polypeptide appears as GARs-(AIRs)2-GARt. However, GARs-AIRs is encoded separately from GARt in yeast, and in bacteria each domain is encoded separately (Henikoff et al., 1997).

1GAR: glycaminamide ribonucleotide synthetase
AIR: aminoimidazole ribonucleotide synthetase
Inferring functional relationships
Domain fusion – Rosetta Stone method

If you find a genome with a fused multidomain protein, and another genome featuring these domains as separate proteins, then these separate domains can be predicted to be functionally linked ("guilt by association")

David Eisenberg, Edward M. Marcotte, Ioannis Xenarios & Todd O. Yeates
Inferring functional relationships

Phylogenetic profiling

If in some genomes, two (or more) proteins co-occur, and in some other genomes they cannot be found, then this joint presence/absence can be taken as evidence for a functional link between these proteins

David Eisenberg, Edward M. Marcotte, Ioannis Xenarios & Todd O. Yeates
Fraction exposed residues against chain length
Fraction exposed residues against chain length
Fraction exposed residues against chain length
Fraction exposed residues against chain length
Fraction exposed residues against chain length
Fraction exposed residues against chain length
Fraction exposed residues against chain length
**Fraction exposed residues against chain length**

If protein structure would be spherical:

- volume is $\frac{4}{3}\pi r^3$
- surface area is $4\pi r^2$

The surface/volume ratio therefore is $3/r$

If a single domain protein grows in size (increasing $r$), the ratio goes down linearly, indicating that the volume increases faster than the surface area.

So, if proteins would just grow by forming larger and larger single domains, then one would expect an increasing fraction of hydrophobic residues (protein core is mostly hydrophobic, surface tends to be hydrophilic).

The plots on the preceding slides show, however, that the fraction of surface (=exposed) residues becomes constant at larger protein sizes (larger numbers of residues), indicating a multi-domain situation.
Analysis of chain hydrophobicity in multidomain proteins
Analysis of chain hydrophobicity in multidomain proteins
Protein domain organisation and chain connectivity

1. β barrel regulatory domain
2. α/β barrel catalytic substrate binding domain
3. α/β nucleotide binding domain

Pyruvate kinase (Phosphotransferase)

Located in red blood cells
Generate energy when insufficient oxygen is present in blood

1 continuous + 2 discontinuous domains
The DEATH Domain

- Present in a variety of Eukaryotic proteins involved with cell death.
- Six helices enclose a tightly packed hydrophobic core.
- Some DEATH domains form homotypic and heterotypic dimers.

http://www.mshri.on.ca/pawson
RGS Protein Superfamily

Founding members of the RGS protein superfamily were discovered in 1996 in a wide spectrum of species

www.unc.edu/~dsiderov/page2.htm
Oligomerisation -- Domain swapping

3D domain swapping definitions. **A:** Closed monomers are comprised of tertiary or secondary structural domains (represented by a circle and square) linked by polypeptide linkers (hinge loops). The interface between domains in the closed monomer is referred to as the C- (closed) interface. Closed monomers may be opened by mildly denaturing conditions or by mutations that destabilize the closed monomer. Open monomers may dimerize by domain swapping. The domain-swapped dimer has two C-interfaces identical to those in the closed monomer, however, each is formed between a domain from one subunit (black) and a domain from the other subunit (gray). The only residues whose conformations significantly differ between the closed and open monomers are in the hinge loop. Domain-swapped dimers that are only metastable (e.g., DT, CD2, RNase **A**) may convert to monomers, as indicated by the backward arrow. **B:** Over time, amino acid substitutions may stabilize an interface that does not exist in the closed monomers. This interface formed between open monomers is referred to as the 0- (open) interface. The 0-interface can involve domains within a single subunit (**I**) and/or between subunits (**II**).
Functional Genomics
Protein Sequence-Structure-Function

We are not good yet at forward inference (red arrows) based on first principles. That is why many widely used methods and techniques search for related entities in databases and perform backward inference (green arrows).

Note: backward inference is based on evolutionary relationships!
Functional Genomics

From gene to function

Genome

Expressome

Proteome

Metabolome
Functional genomics

- The preceding slide shows a simplistic representation of sequence-structure-function relationships: From DNA (Genome) via RNA (Expressome) to Protein (Proteome, i.e. the complete protein repertoire for a given organism). The cellular proteins play a very important part in controlling the cellular networks (metabolic, regulatory, and signalling networks)
Protein structure – the chloroplast skyline

Photosynthesis -- Making oxygen and storing energy in the plant
Protein Function: Metabolic networks controlled by enzymes

Glycolysis and Gluconeogenesis

Proteins are indicated in rectangular boxes using Enzyme Commission (EC) numbers (format: a.b.c.d)
Coiled-coil domains

Tropomyosin

This long protein is involved in muscle contraction