What, Why and How of Computational Protein Structure Prediction

Lukasz Kurgan

Outline

- short and (hopefully) painless introduction to proteins and protein structures
- motivation for computational prediction methods
- overview of computational work in protein structure prediction
- protein structural class prediction

Introduction to Proteins

Brief history

- from the Greek *protas* meaning “of primary importance”
- first proteins were discovered in early 19th century (in 1838) by a Swedish chemist Jöns Jakob Berzelius
  - they were called albuminoids
- for about 100 years chemists argued about their internal structure and finally in 1935 the list of 20 amino acids that compose the proteins was compiled
- nowadays, there are well over 2 millions of known proteins and the detailed structure is known for over 30 thousand of them

Introduction to Proteins

Basic facts

- a protein is a complex, high-molecular-weight organic molecule that consists of amino acids joined by peptide bonds
  - other bio-macromolecules include polysaccharides, lipids, and nucleic acids
- they are among the most actively-studied molecules in biochemistry

Introduction to Proteins

Basic facts

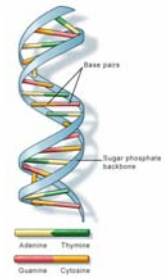
- proteins are essential to the structure and function of all living cells (including humans) and viruses
  - examples functions include catalysis in chemical reactions (enzymes), forming the cytoskeleton (tubulin), serving various signaling and transporting functions (hemoglobin), implementing immune responses (antibodies), regulation of cell processes (hormones), and the list goes on and on...
  - aside from the fat, human body consists of about 20% of proteins by weight
- why are the proteins so “popular”?
  - they can adopt a huge number of three-dimensional shapes and thus constitute a perfect candidate to become a versatile “agent”

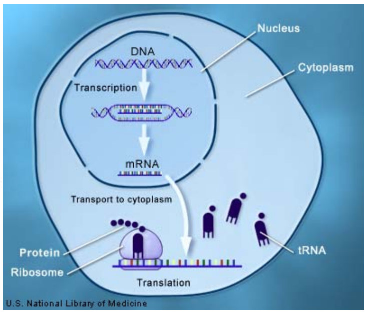
Introduction to Proteins

Who makes the proteins?

- they are assembled from amino acids using information present in genes
  - genes (located in the cell’s nucleus) are transcribed into RNA
  - RNA is then subject to post-transcriptional modification and control, resulting in a mRNA (messanger-RNA) that undergoes translation into a protein
  - mRNA is translated inside a cell by ribosomes that match the three-base codons of the mRNA to the three-base anti-codons of the appropriate tRNA (transfer-RNA)
  - the enzyme aminoacyl tRNA synthetase (aaRs) catalyzes the formation of covalent peptide bonds between amino acids effectively forming a protein chain

Introduction to Proteins





U.S. National Library of Medicine

Introduction to Proteins

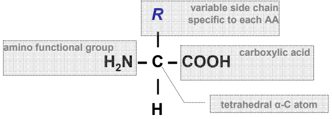
How are they made?

- the translation process produces a linear sequence that is build from amino acids joined by covalent peptide bonds
- nobody really knows (at least for larger proteins) how does it happens that a sequence is transformed into a molecule
  - the sequence folds to form a three dimensional molecule
  - the mechanics of this folding are largely unknown, although we know a lot in terms of the final product of the folding:
    - the molecule can be described on four distinct structural levels
    - it is build from only 20 amino acids
    - and researchers agree that a unique sequence folds always into the same molecule (based on minimum energy principle)

Amino Acids

Amino Acids (AA)

- the basic structural building units of proteins
- they form short polymer chains called peptides or longer polypeptides which are called proteins
- general structure


R
variable side chain specific to each AA
amino functional group
H_2N – C – COOH
carboxylic acid
H
tetrahedral α-C atom

- there are 20 R side chains that make up the different AA

Amino Acids

AA	Abbr.	Side chain	Hydro-phobic	Polar	Electric Charge	Size		Aromatic/ Aliphatic	DNA codon	Occurrence (%)
						Small	Tiny			
Alanine	Ala, A	$-CH_3$	X	-	-	X	X	-	GCU, GCC, GCA, GCG	7.8
Cysteine	Cys, C	$-CH_2SH$	X	-	-	X	-	-	UGU, UGC	1.9
Aspartate	Asp, D	$-CH_2COOH$	-	X	negative	X	-	-	GAU, GAC	5.3
Glutamate	Glu, E	$-CH_2CH_2COOH$	-	X	negative	-	-	-	GAA, GAG	6.3
Phenylalanine	Phe, F	$-CH_2C_6H_5$	X	-	-	-	-	Aromatic	UUU, UUC	3.9
Glycine	Gly, G	-H	X	-	-	X	X	-	GGU, GGC, GGA, GGG	7.2
Histidine	His, H	$-CH_2-C_3H_3N_2$	-	X	positive	-	-	Aromatic	CAU, CAC	2.3
Isoleucine	Ile, I	$-CH(CH_3)CH_2CH_3$	X	-	-	-	-	Aliphatic	AUU, AUC, AUA	5.3
Lysine	Lys, K	$-(CH_2)_4NH_2$	-	X	positive	-	-	-	AAA, AAG	5.9
Leucine	Leu, L	$-CH_2CH(CH_3)_2$	X	-	-	-	-	Aliphatic	UUA, UUG, CUU, CUC, CUA, CUG	9.1
Methionine	Met, M	$-CH_2CH_2SCH_3$	X	-	-	-	-	-	AUG	2.3
Asparagine	Asn, N	$-CH_2CONH_2$	-	X	-	X	-	-	AAU, AAC	4.3
Proline	Pro, P	$-CH_2CH_2CH_2-$	X	-	-	X	-	-	CCU, CCC, CCA, CCG	5.2
Glutamine	Gln, Q	$-CH_2CH_2CONH_2$	-	X	-	-	-	-	CAA, CAG	4.2
Arginine	Arg, R	$-(CH_2)_3NH-C(NH)NH_2$	-	X	positive	-	-	-	CGU, CGC, CGA, CGG, AGA, AGG	5.1
Serine	Ser, S	$-CH_2OH$	-	X	-	X	X	-	UCU, UCC, UCA, UCG, AGU,AGC	6.8
Threonine	Thr, T	$-CH(OH)CH_3$	X	X	-	X	-	-	ACU, ACC, ACA, ACG	5.9
Valine	Val, V	$-CH(CH_3)_2$	X	-	-	X	-	Aliphatic	GUU, GUC, GUA, GUG	6.6
Tryptophan	Trp, W	$-CH_2C_8H_6N$	X	-	-	-	-	Aromatic	UGG	1.4
Tyrosine	Tyr, Y	$-CH_2-C_6H_4OH$	X	X	-	-	-	Aromatic	UAU, UAC	3.2

Protein Structure

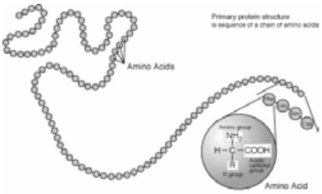
Four distinct aspects of a protein's structure can be defined:

- primary AA sequence
- secondary structure: highly patterned sub-structures of the overall three dimensional structure
  - they include so called α-helices and β-sheets
  - they are defined locally, i.e. many different secondary structure motifs are usually present in a protein molecule
- tertiary structure: the overall shape of a single protein molecule;
  - can be also defined as the spatial relationship of the secondary structural motifs to one another
  - it is primarily formed by hydrophobic interactions; hydrogen bonds, ionic interactions, and disulfide bonds are also involved
- quaternary structure: the shape or structure that results from the union of more than one protein molecule
  - they are called protein subunits, and they function as part of the larger assembly or protein complex.

Protein Sequence

Primary sequence

- proteins are generally relatively large
  - e.g. the muscle protein titin has a 27,000 AA long chain
  - on average about 300 AA

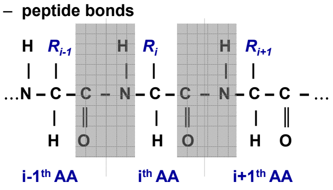


- the two ends of the AA chain are referred to as the carboxy terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity

Protein Sequence

Primary sequence

- peptide bonds



Protein Sequence

Primary sequence

- examples
  - **human hemoglobin 1A3N** (oxygen transporter)

    VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAH
    VDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
  - **immunoglobulin 12E8** (antibody)

    DIVMTQSQKFMSTSVGDRVSITCKASQNVGTAVAWYQQKPGQSPKLMIYSASNRYTGVPDRFTGSGSGTDFT
    LTISNMQSEDLADYFCQQYSSYPLTFGAGTKLELKRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKDINVK
    WKIDGSERQNGVLNSATDQDSKDSTYSMSSTLTLTKDEYERHNSYTCEATHKTSTSPIVKSFNRNEC
  - **narbonin 1NAR** (storage protein)

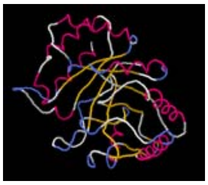
    PKPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPE
    VKVVISIGGRGVNTPFDPAEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITELKKD
    DDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVDYQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTD
    PLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTLQQLLAAR

Protein Sequence

Primary sequence

- examples
  - **narbonin 1NAR** (storage protein)

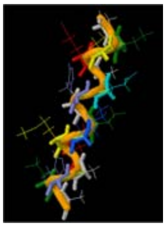
    PKPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPE
    VKVVISIGGRGVNTPFDPAEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITELKKD
    DDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVDYQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTD
    PLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTLQQLLAAR



Protein Structure

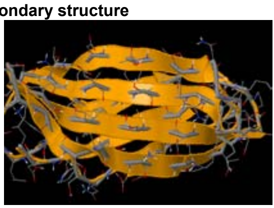
Secondary structure

- **α-helix** structural motif in secondary structure
  - first postulated by Pauling, Corey and Branson in 1951
  - AA are arranged in a helical structure, 5.4 Angstroms (0.54 nanometres) wide
  - all AA side-chains are arranged at the outside of the helix
  - N-H group of nth AA establishes a hydrogen bond with the C=O group of (n+4)th AA
  - on average there are 3.6 AA per turn



Protein Structure

Secondary structure



- **β-strand** structural motif in secondary structure
  - first postulated by Pauling, Corey and Branson in 1951
  - it consists of two or more AA sequences within the same protein that are arranged adjacently and in parallel, but with alternating orientation such that hydrogen bonds can form between the two strands
  - N-H groups in the backbone of one strand establish hydrogen bonds with the C=O groups in the backbone of the adjacent, parallel strand(s)
  - the α-C atoms of adjacent strands are 350 picometres apart

Protein Structure

Secondary structure

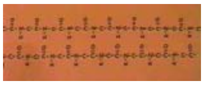
- **β-strand** – a major structural motif in secondary structure
  - **β-sheets** are composed of several β-strands and in general can be classified into two types
    - **parallel β-sheets** where the strands are running in the same direction
    - **anti-parallel β-sheets** where the strands are running in the opposing direction

Protein Structure

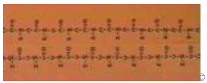
Secondary structure

- α-helix vs. β-sheet
  - protein chain

  - α-helix
  - parallel β-sheet



  - anti-parallel β-sheet



taken from http://dl.clackamas.cc.or.us/ch106-08/secondar.htm

Protein Structure

Secondary structure

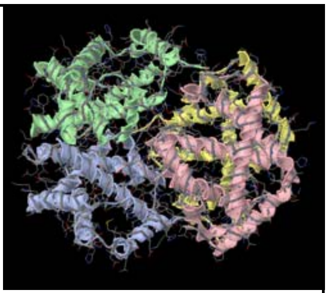
- how is it matched with the primary sequence?
  - DSSP (Dictionary of Protein Secondary Structure) code is frequently used to describe the protein secondary structures for each residue (AA) using a single letter code
    - the secondary structure is assigned based on hydrogen bonding patterns

Protein Structure

Secondary structure

- DSSP codes
  - helices
    - G = 3-turn helix (3/10 helix); min length 3 residues
    - H = 4-turn helix (alpha helix); min length 4 residues
    - I = 5-turn helix (pi helix); min length 5 residues
  - strands
    - E = beta sheet in parallel and/or anti-parallel sheet conformation (extended strand); min length 2 residues
    - B = residue in isolated beta-bridge (single pair beta-sheet hydrogen bond formation)
  - coils (all AA which are not strands or helices)
    - T = hydrogen bonded turn (3, 4 or 5 turn)
    - S = bend (the only non-hydrogen-bond based assignment)

Protein Structure



Secondary structure

- examples
  - human hemoglobin **1A3N** (oxygen transporter)

    different colors denote multiple (4) hemoglobin molecules

primary sequence

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALS
ALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

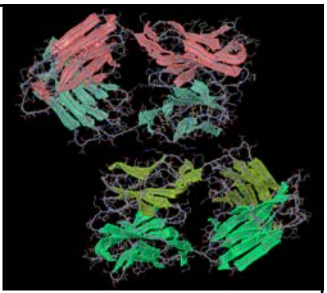
secondary sequence (in 8-states DSSP assignment)

```
CCCHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHHCGGGGGGTTTSCCSTTCHHHHHHHHHHHHHHHHHHHHHTTTSHHHHTH
HHHHHHHHTTCCCTHHHHHHHHHHHHHHHHHHHTTTTTTHHHHHHHHHHHHHHHHHHHTTTCC
```

secondary sequence (in 3-states DSSP assignment)

```
CCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCHHHHHHCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHCCCCHHHHHCH
HHHHHHHHCCCCCCHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHHHHHHHCCCCC
```

Protein Structure



Secondary structure

- examples
  - immunoglobulin 12E8 (antibody)

    different colors denote multiple (4) immunoglobulin molecules

    two dimers are shown

primary sequence

```
DIVMTQSQKFMSTSVGDRVSITCKASQNVGTAVAWYQQKPGQSPKLMIYSASNRYTGVPDRFTGSGSGTDFTLTISNMQSE
DLADYFCQQYSSYPLTFGAGTKLELKRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNS
ATDQDSKDSTYSMSSTLTLTKDEYERHNSYTCEATHKTSTSPIVKSFNRNEC
```

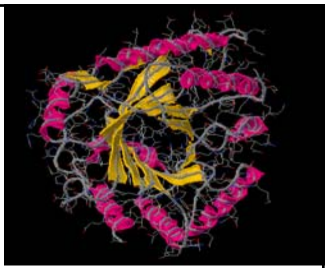
secondary sequence (in 8-states DSSP assignment)

```
CCCEECCCSEEEECTTCCEEEEEEEESSCCTTCEEEEEECTTSCCEECEETTTEECTTTTTTEEEEEETTEEEEEEESSCCGG
GCSEEEEEECSSSSCEECCCEEEEECCCCBCCEEEEECCCHHHHTTTEEEEEEEEEESEESSCCEEEEEEETTEECCTTEEEE
ECCCCTTTCCBCEEEEEEEEEHHHHTTCSEEEEEEECTTCSSCEEEEEEETTTT
```

secondary sequence (in 3-states DSSP assignment)

```
CCCEECCCCEEEECCCCCEEEEEEEECCCCCCCEEEEEECCCCCCEECEECCCEECCCCCCCEEEEEECCEEEEEEECCCCHH
HCCEEEEEEECCCCCCEECCCEEEEECCCCECCEEEEECCCHHHHCCCEEEEEEEEEECEECCCCEEEEEEECCEECCCCEEEE
ECCCCCCCCCECEEEEEEEEEHHHHCCCCEEEEEEECCCCCCCEEEEEEECCCC
```

Protein Structure



Secondary structure

- examples
  - narbonin 1NAR (storage protein)

    one molecule is shown

primary sequence

```
KPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVI
SIGGRGVNTPFDPAEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITELKKDDDLNINVVS
IAPSENNSSHYQKLYNAKKDYINWVDYQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTDPLDTKHNKITRDIFIG
GCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTLQQLLAAR
```

secondary sequence (in 8-states DSSP assignment)

```
CCEEEEEEESCCTTCCSCSSCCSTTCCCSSEEEEEEEECCEEECBCTTSCBCSCEEECSCHHHHTHHHHHHHHHCTTCEEE
EEEEESSTTSCBCBSCTTTHHHHHHHHHHHHHHSEETTEECCCEEEEEESCBCSSTTHHHHHHHHHHHHHHCTTSCCCEE
EECCCTTTHHHHHHHHHTTTTCCEEEEEGGGCSSCCCSHHHHHHHHHHHHHHSCTTCEEEEEECCHHHHHHCSSCHHHHH
HHHHHHHHTTCCCEEEEECHHHHSSCSSTTTTTTHHHHHHHHHHHCC
```

secondary sequence (in 3-states DSSP assignment)

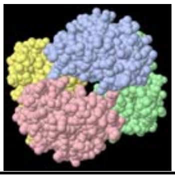
```
CCEEEEEEECCCCCCCCCCCCCCCCCCCCEEEEEEECCEEECECCCCCECCCEEECCCHHHHCHHHHHHHHHHCCCCEEE
EEEEECCCCCCECECCCCCHHHHHHHHHHHHHHHCEECCEECCCEEEEEECCECCCCCHHHHHHHHHHHHHHCCCCCCCEE
EECCCCCCHHHHHHHHHHCCCCCCEEEEEHHHCCCCCCCHHHHHHHHHHHHHHCCCCCEEEEEECCHHHHHHCCCCHHHHH
HHHHHHHHCCCCCEEEEECHHHHCCCCCCCCCCCHHHHHHHHHHHCC
```

Protein Structure

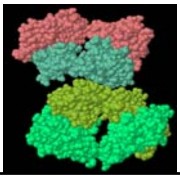
Tertiary and Quaternary structures

- the overall shape of a single (or a multi) protein molecule
  - a spatial arrangement of the secondary structural motifs
  - formed by hydrophobic interactions, hydrogen bonds, ionic interactions and disulfide bonds

human hemoglobin
colors show individual proteins



immunoglobulin
colors show individual proteins



narbonin
colors shown secondary structures

Protein Structure

Why is it important to know the structure?

- in short: knowing the structure allows us to modify, e.g. enhance or block, certain protein functions
  - example: if a protein is involved in cell division and we block this function, we effectively stop the cell from dividing
    - when would that be useful?

Protein Structure

Why is it important to know the structure?

- various molecules/ions can bind to specific protein sites
  - the sites are called binding sites and exhibit chemical specificity
  - the particle that binds is called a ligand
  - the strength of ligand-protein binding is a property of the binding site known as affinity
- since proteins are involved in practically every function performed by a cell, the mechanisms for controlling these functions therefore depend on controlling protein activity
  - regulation can involve a changing protein's shape or concentration, e.g.:
    - allosteric modulation: binding of a ligand at one site on a protein affects the binding of a ligand at another site
    - covalent modulation: covalent modification of a protein affects the binding of a ligand or some other aspect of the protein's function

Protein Structure

How do we learn the sequence?

- can be deduced from known DNA sequence
- can be learned based on Edman degradation and mass spectrometry methods

- relatively cheap and easy to perform for virtually all proteins

Protein Structure

How to we learn the structure?

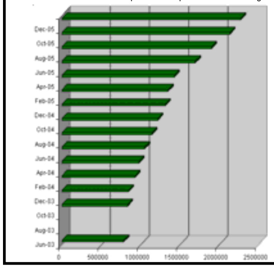
- tertiary structures are deduced through crystallography or multidimensional NMR
  - secondary structure is computed from the tertiary structure
  - crystallography = X-ray of a crystallized protein
  - multidimensional NMR = Nuclear Magnetic Resonance Spectroscopy of aqueous samples of highly purified protein
    - uses magnetic properties of a nuclei
  - problems
    - costly and labor extensive
    - some proteins cannot be crystallized or purified

Experimental method	X-ray	27531
	NMR	4436
	Electron Microscopy	79
	Other	70
	Total	32116

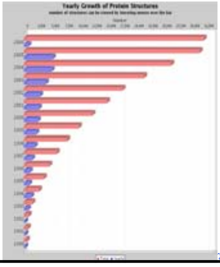
The Gap

So, what is the problem?

number of known proteins
based on NCBI Reference Sequences at http://www.ncbi.nlm.nih.gov/RefSeq/



number of protein for which (tertiary) structure is known
based on Protein Data Bank at http://www.rcsb.org/pdb



Protein Databases

Databases

- PDB
  - the single worldwide repository for the processing and distribution of 3-D structure of proteins
  - manually curated and annotated (by experts) database of known tertiary protein structures
  - URL: http://www.rcsb.org/pdb

Protein Databases

Databases

- SWISS-PROT
  - protein knowledgebase established in 1986 and maintained since 2003 by the UniProt Consortium
    - a collaboration between the Swiss Institute of Bioinformatics and the Department of Bioinformatics and Structural Biology of the Geneva University, the European Bioinformatics Institute (EBI) and the Georgetown University Medical Center's Protein Information Resource (PIR)
  - manually annotated and curated (by experts) database of known protein sequences and relevant information
    - protein function, domain structure (if known), variants, post-translational modifications, similarities to other proteins
  - URL: http://ca.expasy.org/sprot/sprot_details.html

Protein Databases

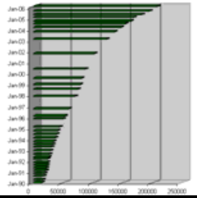
Databases

- NCBI
  - integrated access to a variety of sources, including SwissProt, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq
  - proteins are submitted and managed by individual researchers
    - they are not curated by experts
  - contains mainly protein sequences and relatively little additional information
  - URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=protein

The Gap

So, what is the problem here?

• # of known proteins (NCBI):	2,273,764 (January, 2006)
• # of proteins for which structure is known (PDB):	32,116 (January 2006)
• # of protein for which high quality information (i.e. sequences, partial secondary structure, etc.) is known (SWISS PROT):	207,132 (February 2006)



The Gap

What can we do to close the GAP?

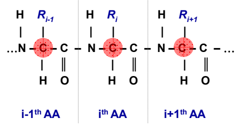
- develop computational method to predict the structure based on the available information
  - mainly the primary sequence is used, but we could also use protein function and other known information
    - see SWISS-PROT
  - computational = cheap and can work without the restrictions enforced by experimental conditions

….so far the quality of the computational methods is not sufficiently good, but it is constantly improving

Protein Structure Prediction

Computational prediction

- the **ultimate goal** is to predict the native conformation of a protein from its primary sequence
  - the prediction boils down to spatial placement of the central α-C atoms for each AA in the sequence

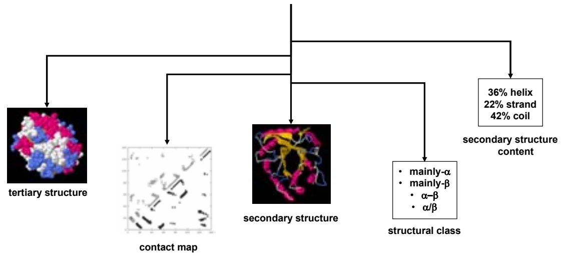


  - since direct methods are not successful, a number of other prediction methods is researched

Protein Structure Prediction

INPUT: primary sequence (narbonin 1NAR)

```
KPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVISIGGRGVNTPFDPAEENVWVSNA
KESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITELKKDDDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVDYQFSNQQKPVSTDDAFVEIF
KSLEKDYHPHKVLPGFSTDPLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTLQQLLAAR
```



tertiary structure

contact map

secondary structure

- mainly-α
- mainly-β
  - α–β
  - α/β

structural class

36% helix
22% strand
42% coil

secondary structure content

Protein Structure Prediction

NARBONIN (1NAR) protein

PKPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVISIGGRGVNTPFDPAEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITE
LKKDDDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVDYQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTDPLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTLQQLLAAR

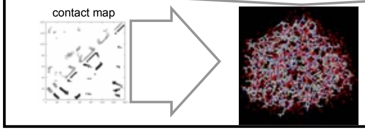
Molecular Weight	AA COMPOSITION	RESIDUE PROPERTIES (electric charge, chemical group, etc.)	HYDROPHOBICITY based residue properties
33071.5	0.0345 0.0034 0.0759 0.0724 0.0586 0.0586 0.031 0.0828 0.0793 0.0759 0.0034 ….	0.3207 0.1103 0.2793 0.1414 0.1483 0.2793 0.2379 0.2276 0.4931 0.2172 0.2414 0.2241 0.169 0.1897 0.1724 ….	0.2544 0.1154 0.2198 0.2893 0.1343 -0.02268 138.8 0.4788 139.6 0.05426 -0.05693 0.01952 0.03342 ….

secondary structure content prediction
helix 35.9%; strand 21.7%; coil 42.4%

structural class classification
α-β proteins

secondary structure classification (black coil, yellow strand, red helix)

PKPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVISIGGRGVNTPFDPAEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITE
LKKDDDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVDYQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTDPLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTLQQLLAAR



contact map

Protein Structure Prediction

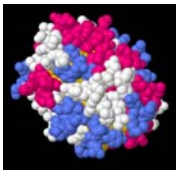
Main computational prediction tasks

- overall tertiary structure

INPUT: primary sequence (narbonin 1NAR)

```
KPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVISIGGRGVNTPFDP
AEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITELKKDDDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVD
YQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTDPLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTL
QQLLAAR
```

OUTPUT: tertiary structure



Protein Structure Prediction

Main computational prediction tasks

- overall tertiary structure
  - **de novo** (ab initio) methods
    - they build tertiary protein models "from scratch“
    - they attempt to mimic protein folding or apply some stochastic method to search possible solutions (i.e. global optimization of a suitable energy function)
    - they require vast computational resources (supercomputers, such as Blue Gene or distributed computing) and thus have only been carried out for very small proteins

Protein Structure Prediction

Main computational prediction tasks

- overall tertiary structure
  - **comparative** methods
    - they use previously solved structures (or templates) as starting point
      - » research shows that there are only around 2000 distinct protein folds in nature, though there are many millions of different proteins
    - **homology-based** methods are based on an assumption that two homologous proteins (proteins with similar sequence) will share very similar structures
    - **threading-based** methods scan sequence of an unknown structure against a database of solved structures and a scoring function is used to assess the compatibility of the sequence to the structure, thus yielding possible three-dimensional models

Protein Structure Prediction

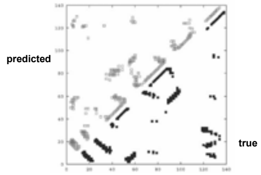
Main computational prediction tasks

- contact maps

INPUT: primary sequence (narbonin 1NAR)

```
KPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVISIGGRGVNTPFDP
AEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITELKKDDDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVD
YQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTDPLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTL
QQLLAAR
```

OUTPUT: contact map



Protein Structure Prediction

Main computational prediction tasks

- contact maps
  - tertiary structure of a protein can be captured to a large extent by its distance map
    - the distance map is a two-dimensional symmetric matrix that shows which tuples of protein elements are close to each other in the overall molecule
      - » elements range between atoms, through AAs, to segments of secondary structure
      - » in case of amino-acids distances are usually calculated between α-C atoms

Protein Structure Prediction

Main computational prediction tasks

- overall secondary structure

INPUT: primary sequence (narbonin 1NAR)

```
KPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVISIGGRGVNTPFDP
AEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITELKKDDDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVD
YQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTDPLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTL
QQLLAAR
```

OUTPUT: secondary structure

```
CCEEEEEEECCCCCCCCCCCCCCCCCCCCCEEEEEEECCEEECECCCCCECCCEEECCCHHHHCHHHHHHHHHHCCCCEEEEEEEECCCCCCEC
ECCCCCHHHHHHHHHHHHHHHCEECCEECCCEEEEEECCECCCCCHHHHHHHHHHHHHHCCCCCCCEEEECCCCCCHHHHHHHHHCCCCCCEE
EEEHHHCCCCCCCHHHHHHHHHHHHHHCCCCCEEEEEECCHHHHHHCCCCHHHHHHHHHHHHHHCCCCCEEEEECHHHHCCCCCCCCCCCHHHHH
HHHHHHCC
```

Protein Structure Prediction

Main computational prediction tasks

- overall secondary structure
  - comparative methods (dominant)
    - they use previously solved structures (or templates) as starting point
    - for the predicted sequence a search for known homologous sequence is used and the structure is inferred based on the structure of these sequences

Protein Structure Prediction

Main computational prediction tasks

- secondary structure content

INPUT: primary sequence (narbonin 1NAR)

```
KPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVISIGGRGVNTPFDP
AEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITELKKDDDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVD
YQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTDPLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTL
QQLLAAR
```

TRANSFORMED INPUT: feature space representation of the sequence

Molecular Weight	AA COMPOSITION	RESIDUE PROPERTIES (electric charge, chemical group, etc.)	HYDROPHOBICITY based residue properties
33071.5	0.0345 0.0034 0.0759 0.0724 0.0586 0.0586 0.031 0.0828 0.0793 0.0759 0.0034 ….	0.3207 0.1103 0.2793 0.1414 0.1483 0.2793 0.2379 0.2276 0.4931 0.2172 0.2414 0.2241 0.169 0.1897 0.1724 ….	0.2544 0.1154 0.2198 0.2893 0.1343 -0.02268 138.8 0.4788 139.6 0.05426 -0.05693 0.01952 0.03342 ….

OUTPUT: secondary structure content

helix 0.359, strand 0.217, coil 0.424

Protein Structure Prediction

Main computational prediction tasks

- secondary structure content
  - classical prediction uses a feature-based representation of a sequence as the input
  - percentage amount of each of the three main secondary structures (helices, strands and coils is predicted)
    - either neural networks or multiple-regression methods are used

Protein Structure Prediction

Main computational prediction tasks

- secondary structure content (and structural class – next)
  - classical prediction uses a feature-based representation of a sequence as the input
    - protein sequence length
    - molecular weight
    - isoelectric point
    - composition vector
    - composition moment vector
    - dipeptide composition
    - AA groups: R-groups, exchange groups, hydrophobicity groups, electronic groups, chemical groups, other groups
    - for details see

      Kurgan L and Homaeian L, Prediction of Secondary Protein Structure Content from Primary Sequence Alone - a Feature Selection Based Approach, *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition* (MLDM 2005), pp. 334-345, 2005

Protein Structure Prediction

Main computational prediction tasks

- structural class

INPUT: primary sequence (narbonin 1NAR)

```
KPIFREYIGVKPNSTTLHDFPTEIINTETLEFHYILGFAIESYYESGKGTGTFEESWDVELFGPEKVKNLKRRHPEVKVVISIGGRGVNTPFDP
AEENVWVSNAKESLKLIIQKYSDDSGNLIDGIDIHYEHIRSDEPFATLMGQLITELKKDDDLNINVVSIAPSENNSSHYQKLYNAKKDYINWVD
YQFSNQQKPVSTDDAFVEIFKSLEKDYHPHKVLPGFSTDPLDTKHNKITRDIFIGGCTRLVQTFSLPGVFFWNANDSVIPKRDGDKPFIVELTL
QQLLAAR
```

TRANSFORMED INPUT: feature space representation of the sequence

Molecular Weight	AA COMPOSITION	RESIDUE PROPERTIES (electric charge, chemical group, etc.)	HYDROPHOBICITY based residue properties
33071.5	0.0345 0.0034 0.0759 0.0724 0.0586 0.0586 0.031 0.0828 0.0793 0.0759 0.0034 ….	0.3207 0.1103 0.2793 0.1414 0.1483 0.2793 0.2379 0.2276 0.4931 0.2172 0.2414 0.2241 0.169 0.1897 0.1724 ….	0.2544 0.1154 0.2198 0.2893 0.1343 -0.02268 138.8 0.4788 139.6 0.05426 -0.05693 0.01952 0.03342 ….

OUTPUT: structural class

α/β class

Protein Structure Prediction

Main computational prediction tasks

- structural class
  - a number of definitions of a structural class were developed

reference	structural class	helix (α) amount	strand (β) amount	additional constrains and comments
Nakashima et al., 1986	α proteins β proteins α+β proteins α/β proteins irregular	> 15% < 15% > 15% > 15%	< 10% > 10% > 10% > 10%	contains dominantly antiparallel β-sheets contains dominantly parallel β-sheets otherwise
Chou, 1995	α proteins β proteins α+β proteins α/β proteins ξ proteins	≥ 40% ≤ 5% ≥ 15% ≥ 15% ≤ 10%	≤ 5% ≥ 40% ≥ 15% ≥ 15% ≤ 10%	more than 60% antiparallel β-sheets more than 60% parallel β-sheets
Eisenhaber et al., 1996	α proteins β proteins mixed proteins irregular	> 15% < 15% > 15%	< 10% > 10% > 10%	otherwise
SCOP Murzin et al., 1995	α proteins β proteins α+β proteins α/β proteins + 7 other classes	N/A	N/A	manual classification

Protein Structure Prediction

Main computational prediction tasks

- structural class
  - Structural Classification of Proteins (SCOP) structural classes
    - URL: http://scop.mrc-lmb.cam.ac.uk/scop/
    - does not incorporate hard-coded rules for structural classes
    - classification is manual based on structural elements that are located in individual domains that constitute the protein
    - includes eleven classes: 1) all-α proteins; 2) all-β proteins; 3) α/ β proteins; 4) α+ β proteins; 5) multi-domain proteins; 6) membrane and cell surface proteins; 7) small proteins; 8) coiled coils proteins; 9) low resolutions proteins; 10) peptides; and 11) designed proteins
    - usually, only the first four categories are considered for computational prediction purposes as they include significant majority of the protein sequences

Structural Class Prediction

State-of-the-art

- "chaotic progress"
  - over a dozen prediction methods, which were never comprehensively compared, were proposed
  - very basic protein representation
    - composition vector + polypeptide composition
  - no established test beds
    - each method tested on a different datasets
    - variable homology
    - "cheating"
    - test types: resubstitution and jackknife

Structural Class Prediction

sequence representation

size and homology of test datasets

cheating in design and testing

classification algorithm	represen-tation	classes	dataset				classification accuracy		
			size	homology	domains	reference	resub	jackknife	reference
Vector decomposition	AA compos. vector	3 classes Eisenhaber et al., 1996	260	unknown	no	Eisenhaber et al., 1996	60.8	57.7	Eisenhaber et al., 1996
			471	unknown	no		58.2	57.3	
Geometric classification	AA comp vector	4 classes SCOP	359	unkn, but homologous	yes	Chou & Maggiora, 1998	94.3	84.1	Chou & Maggiora, 1998
Component coupled geometric classification	AA comp vector	4 classes SCOP	359	unkn, but homologous	yes		94.4	84.7	Bu at al., 1999
	energy auto-correlation functions	4 classes SCOP	359	unkn, but homologous	yes		96.7	90.5	
Bayes classification	AA comp vector	4 classes Nakashima et al., 1986	131	unknown	no	Nakashima et al., 1986	99.2	42.7	Wang & Yuan, 2000
		4 classes Chou, 1995	120	unknown	no	Chou, 1995	100	53.3	
		3 classes Eisenhaber et al., 1996	260	unknown	no	Eisenhaber et al., 1996	86.5	62.7	
			471	unknown	no		79.6	66.7	
		4 classes SCOP	1189	40%	yes	Wang & Yuan, 2000	63.8	***53.8***	
			675	30%	yes		66.7	***48.0***	
Discriminant analysis	AA&polypeptide comp vector	4 classes SCOP	1054	40%	yes	Luo et al., 2002	91.7	75.2	Luo et al., 2002
	AA comp vector	4 classes SCOP	1054	40%	yes		66.2	***55.8***	
Information discrepancy based classification	polypeptides	4 classes SCOP	359	unkn, but homologous	yes	Chou & Maggiora, 1998	---	95.8	Jin et al., 2003
		4 classes SCOP	1401	30%	yes	Jin et al., 2003	---	75.0	
Support Vector Machines	AA comp vector	4 classes SCOP	359	unkn, but homologous	yes	Chou & Maggiora, 1998	93.0	95.2	Cai et al., 2003
	AA comp vector	7 classes SCOP	1601	unkn, but homologous	yes		87.0	84.1	
Intimate sorting classification	AA comp vector functional domain composition	7 classes SCOP	2230	20%	yes	Chou & Cai, 2004	---	98.8	Chou & Cai, 2004

Structural Class Prediction

Our study

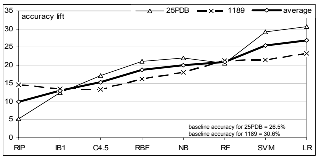
- multi-goal study, which includes investigation of eight prediction algorithms
  - Naïve Bayes (NB), Radial Basis Function neural network (RBF), Instance Based classifier (IB1), C4.5 (C4.5), Random Forest (RF), Repeated Incremental Pruning to Produce Error Reduction (RIP), Support Vector Machine (SVM), and Logistic Regression (LR)
- three datasets with different homologies
  - two low homology (1189 and 2340 sequences) and one high homology (359 sequences)
- three protein sequence representations
  - 1) composition vector, 2) energy autocorrelation, 3) newly proposed representation based on composition and composition moment vectors vector, chemical group composition, hydrophobic autocorrelations and molecular weight
- and finally three test procedures
  - resubstitution, jackknife, 10-fold-cross-validation

Structural Class Prediction

Prediction (classification) algorithms

- hard problem 50+% accuracy
- best are SVMs and logistic regression

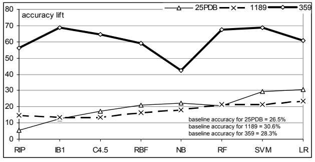
classification algorithm (1189 dataset)	representation	classification accuracy		
		resubsti -tution	jackk nife	reference
Support Vector Machine	AA composition vector	57.8	52.3	this paper (2nd best result)
Bayes classification	AA composition vector	63.8	53.8	Wang and Yuan, 2000
Logistic regression	66 features	62.0	53.9	this paper (best result)



Structural Class Prediction

Sequence homology

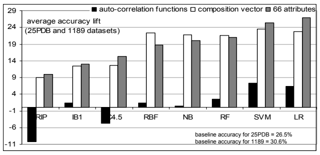
- a paired t-test between the results achieved by the eight algorithms on the 25PDB and 359 datasets gave t-score of 10.0 and between the 1189 dataset and 359 dataset gave t-score of 13.0
  - the difference is statistically significant
- a paired t-test between the results for the 25PDB and 1189 datasets resulted in t-score of 1.0
  - the difference is statistically not significant



Structural Class Prediction

Sequence representation

- high quality of the composition vector with respect to structural class prediction was confirmed
- increase of accuracy lift due to using the new representation is 2.0% for support vector machines and 4.3% for logistic regression
  - the improvements concern the most accurate classifiers



Structural Class Prediction

Test procedures

- resubstitution test is unreliable and should not be reported
- 10-fold cross-validation is shown to be at least the same good as the jackknife test
  - execution of the jackknife test requires substantial computational time, in comparison with less demanding and commonly performed 10-fold cross-validation (evaluation of the logistic regression method using 10-fold cross-validation requires about 50 minutes, and using jackknife test about 8400 minutes)

	dataset	25PDB			1189			359		
	representation	CV	AC	66	CV	AC	66	CV	AC	66
10 fold cross-validation compared with jackknife	t-test result	=	=	=	--	=	=	--	=	--
	t-score	0.1	1.3	1.1	2.5	0.9	0.8	3.0	1.8	3.8
	confidence level	N/A	N/A	N/A	> 97%	N/A	N/A	> 99%	N/A	>99.5%

Structural Class Prediction

The study have shown that

- sequence homology is found to significantly affect accuracy
- new to the field logistic regression prediction algorithm generates results that are competitive or better when compared with the past results
- for eight considered prediction algorithms, state-of-the-art sequences representation and low, about 30%, homologous dataset, the best results are in the range of 57% accuracy
- the newly proposed sequence representation is beneficial for high quality prediction algorithms
- the resubstitution tests are shown to significantly overestimate the prediction accuracy, and the commonly performed jackknife test procedure leads to unnecessarily high computational demand
  - therefore 10-fold cross-validation should be used for the future studies

Protein Structure

Where to get started?

- tertiary structure
  - overall tertiary structure

    Kolinski A, Protein modeling and structure prediction with a reduced representation, *Acta Biochim Pol.*, 51(2), pp.349-71, 2004

    Bujnicki JM, Protein-structure prediction by recombination of fragments, *Chembiochem*, 7(1), pp.19-27, 2006
  - contact maps

    Pollastri G and Baldi P, Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners, *Bioinformatics*, 18, Suppl 1:S62-70, 2002

    MacCallum RM, Striped sheets and protein contact prediction, *Bioinformatics*, 20, Suppl 1, I224-I231, 2004

Protein Structure

Where to get started?

- **secondary structure**
  - **overall secondary structure**

    Heringa J, Computational methods for protein secondary structure prediction using multiple sequence alignments, *Curr Protein Pept Sci.*, 1(3), pp.273-301, 2000

    Przybylski D and Rost B, Alignments grow, secondary structure prediction improves, *Proteins*, 6(2), pp.197-205, 2002
  - **structural class**

    Chou KC, Progress in protein structural class prediction and its impact to bioinformatics and proteomics, *Curr Protein Pept Sci.*, 6(5), pp.423-36, 2005

    Wang Z-X, and Yuan Z, How Good is the Prediction of Protein Structural Class by the Component-Coupled Method?, *Proteins*, 38, pp.165-175, 2000
  - **secondary structure content**

    Lee S, Lee BC and Kim D, Prediction of protein secondary structure content using amino acid composition and evolutionary information, *Proteins*, 62(4), pp.1107-14, 2006

    Lin Z and Pan XM, Accurate prediction of protein secondary structural content, *J Protein Chem*, 20(3), pp.217-20, 2001