Integrated Structural Bioinformatics

Áron Erdélyi

2021.09.22.

Contents

1 The added value of structural informatics

1.1 Understanding biomolecular function is only possible if we know the structure

The best known example is the DNA double helix. This explained, how genetic information is stored, which was an (almost) impossible task, if we don't know the structure. It also proposed a mechanism of copying of the genetic material explaining inheritance. Today we use this knowledge is used to sequence DNA.

Before solving the structure of the ribosome, we didn't know, that the peptide bond is catalyzed by RNA. The structure also helped to understand the mechanisms underlying codon recognition, including wobbling, and the mechanism of action of a number of antibiotics.

Virus protease structures are invaluable help in designing antiviral drugs.

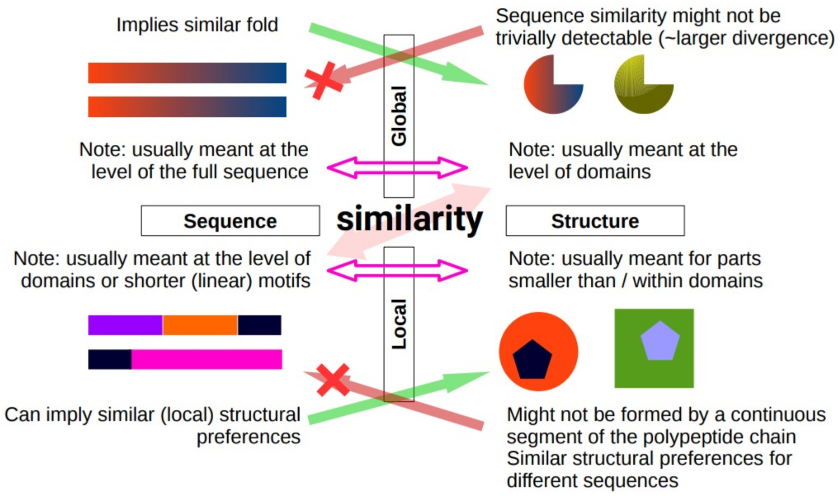
1.2 What is structural bioinformatics

Structural bioinformatics is the science that helps in the assessment, interpretation and making use of structural features in biomolecules.

- Assessing what the given structure can be used for
  - Representation of biomolecular structures
  - Quality assessment of structures.
- Interpretation of the structure based on local and global similarities
  - Identifying structural features
    - Secondary structures
    - Structural domains
  - Assessing function from structure
- Generating structure–related information
  - Prediction of structural features
    - Assignment of a property to each residue
    - Prediction of full 3D structures
  - Design of structures with desired properties and functions.

1.3 Similarity is a key concept in bioinformatics

Global similarity implies evolutionary relationships, while local similarity does not necessarily imply this relationship, but it is usually because of a functional relevance.


Implies similar fold
Sequence similarity might not be trivially detectable (~larger divergence)
Global
Note: usually meant at the level of the full sequence
Note: usually meant at the level of domains
Sequence
similarity
Structure
Note: usually meant at the level of domains or shorter (linear) motifs
Note: usually meant for parts smaller than / within domains
Local
Can imply similar (local) structural preferences
Might not be formed by a continuous segment of the polypeptide chain
Similar structural preferences for different sequences

2 Representation of molecules and macromolecules, the PDB database

2.1 Models in bioinformatics

In bioinformatics we have four kinds of data, that we store in databases:

- Sequences
- 3D
- Networks
- Text

2.1.1 Generalized structures as database records

A database record consists of two main parts, annotation and the actual data. Annotation labels can be on of the following types:

- **Global descriptors**, for example the source of the data, biological role and function, etc...
- **Local descriptors**, for example binding sites and domains.

2.2 Representation of molecules and chemical substances

We can represent molecules with different approaches. These include the molecular surface, 2D formula, SMILES, InChl, ect...

To represent chemical substances, we must take into consideration the following key aspects.

- Information content (exact description)
- Extendability
- Unabiguity and uniqueness
- Reproducibility
- Feasibility
- Availability of programs that support the given format for display, manipulation, calculations etc.
- Availability of converters that support the format

2.3 Description of the 3D structure of molecules

When modelling molecules one must represent points in 3D space. It is obvious that for this we use a 3D Cartesian coordinate system, but the question arises, what unit should we use. The obvious answer in many cases is the use of Angstrom as a unit.

The other question is how to tell if two molecules are connected. We should take into consideration if the format tells explicitly if they are connected or not.

It is common for programs to use an internal coordinate system. This includes bond lengths, bond angles and torsion angles, meaning for N atoms, we will have $3N - 6$ parameters.

2.4 Representing protein structures

Representing protein structures is a bigger problem that representing simple molecules. We must decide the level of representation.

Representation of the primary level is quite easy, and it is also necessary. We must know the coordinates of all atoms and almost all bonds. This is typically represented as a string.

Representing secondary and higher level structures is not such a trivial problem. It is done by reduced representation.

We can use topology diagrams in 2D to represent secondary structures. The elements include:

- Their position in the sequence
- Some of their key connections

This 2D representation can be useful, when comparing structures in some cases.

2.5 The Protein Data Bank – PDB

2.5.1 The database

The PDB is the primary protein structure database. It contains structures of proteins, nucleic acids and their complexes. The source of the data is X-ray, NMR, and other atomic resolution structure determination experiments.

Structures are stored in the database as 3D atomic coordinates.

2.5.2 The file format

The PDF file format was developed for biological molecules and it is the most widely used format for this purpose. It is a very strict format.

The file consists of two parts. The header and the atomic coordinates. By default there is no connectivity information. Programs either know the topology or guess bonds from atomic distances.

As the length of each data field is limited, it is unsuitable for large macromolecular complexes.

- Numbering of amino acid residues can be highly chaotic due to efforts to conform the consensus numbering in some protein families.
- In X-ray structures, coordinates of highly mobile atoms might often be missing.
- The last column does not always contain the chemical symbol of the atom although it is required by some programs.
- Chain are terminated by TER records, and the file ends with an END record.
- In crystal structures multiple well-defined conformers may be present for some residues, these can be given explicitly by their coordinates using fractional occupancy values (their sum should be 1, however)
- For structures determined by NMR, usually more conformers are deposited. Conformers are separated by MODEL/ENDMDL keywords. This also provides a way to overcome some of the difficulties of storing large multichain structures, as the same chain ID can be used for each structure.
- There are many programs that read/write PDB format files. Problem is, they often use their own nomenclaure for hydrogen atoms and sometimes also for other atoms like terminal oxygens etc. This greatly adds to the chaos, mainly affecting NMR spectroscopists and people doing modeling/molecular dynamics, who explicitly deal with hydrogen atoms.
- The PDB format is a highly restricted format hardly appropriate to store information on large complexes determined in the last decade (e.g. virus capsids or the ribosome) - strange chain identifiers etc.
- Due to their complexity structural data are among the hardest ones to check for errors
- Reliability can vary within parts of an entry in a way that is hard to judge

Some other formats include PDBx/mmCIF and PDBML.

3 Validation of macromolecular structures

All protein structures are only models. However, in structural biology the term model is commonly used for structures prepared without experimental data.

3.1 Structure determination methods

3.1.1 Low resolution methods

Low resolution methods are used in secondary structure determination. Used methods include circular dichorism, FT-IR spectroscopy, ect...

3.1.2 At atomic resolution

The three most commonly used methods are

- **X-ray crystalography**: the method of choice for most proteins, enormous advances in the investigation of large supramolecular complexes in the past two decades.
  - Molecule should be crystalized
  - Well-difracting crystals measured almost exclusively at synchrotons (powerful X-ray source)
  - Result: diffraction pattern with information on
    - Crystal type
    - Molecular structure
  - Obstacle: Heavy phase problem (only intensity info recorded)
    - Heavy atom incorporation
    - Homologous replacement
    - MAD (multiwavelength anomalous diffraction)
    - Approximately correct initial phase can then be refined
  - If phase is OK, a correct electron density map can be obtained
  - Deposited: single, or a limited number of conformers
  - In crystallography, structures are fit into the electron density map. Sometimes multiple well defined conformers may be present for some residues, these can be given explicitly by their coordinates using fractional values in the PDB file (their sum should be one).
- **NMR spectroscopy**: typically used when crystalization is not feasable, like for small proteins, but its largest potential lies in the investigation of internal dynamics, key advances in this aspect in the last two decades.
  - Parameters with 3D structural information:
    - Chemical shifts
    - Nuclear Overhauser Effect (H-H distances, less than 6 Angstrom)
    - Relative bond orientations
  - Routine methods: find single conformers fitting all the data
  - Usually a number of similar conformers fulfill the parameters
  - Claim that these reflect the internal dynamics of the molecule in the solution cannot be justified for conventional structure calculation methods (all conformers forced to fulfill all parameters)
  - More recent methods consider ensemble averaging
  - Deposited: conformer set or minimized average
  - Every peak corresponds to two spatially close atoms. The task is to identify these atoms and use the distances in the structure calculation while iteratively correcting errors.
  - For structures determined by NMR, usually more conformers are deposited. Most of the time in PDB files these conformers are seperated by MODEL/ENDMDL keywords.
  - Conventional methods mainly use NOE-based (H-H) distances.

  - Solution: e.g. simulated annealing scheme (heating the molecule up and then gradually cooling it down) in molecular dynamics with distances as restraints
  - Iterative refinement: correcting errors, gathering more information
  - Today automated methods are available: yield a good fold, but throw away a large portion of the data

- **Cryo-electron microscopy**: reached atomic resolution in the past few years.
  - Errors of atomic positions and B-factos:
    - Theoretically possible, but enormous computing capacity is needed (matrix inversion)
    - Available only for several small proteins in the PDB
  - Resolution:
    - Denotes the distance between crystallographic planes for which reflections are available. Not unified: either best layer with a full set of reflections or the best available reflections considered.
    - Greater than 4A: only molecular shape
    - 3A: main chain
    - 2A: side chains
    - 1.2-0.9A: hydrogen atoms, alternate conformers visible.
  - R-factor:
    - Difference between actual and back-calculated diffraction pattern (OK bellow 20%)
    - Free R-factor: leaving 5% of experimental data out for structure calculations, checking the above difference for this part, 40% is OK.

3.2 Quality measures for determination methods

3.2.1 X-ray crystallography

- B-factors (for each atom):
  - Reflects the uncertainty of atomic positions
  - Influenced e.g. by thermal motions

3.3 NMR spectroscopy

- NOE violations:
  - Should not exceed 0.5A.
  - Number of NOE peaks per resedue yields a rough estimate of quality
- RMSD (Root mean square deviation - of atomic positions after superimposing all structures)
  - Considered OK around 1A.
  - Not unified
- Efforts to create R-factor like measures by calculating back spectra/chemical shifts etc.
- Rule of thumb based on the early NMR structures

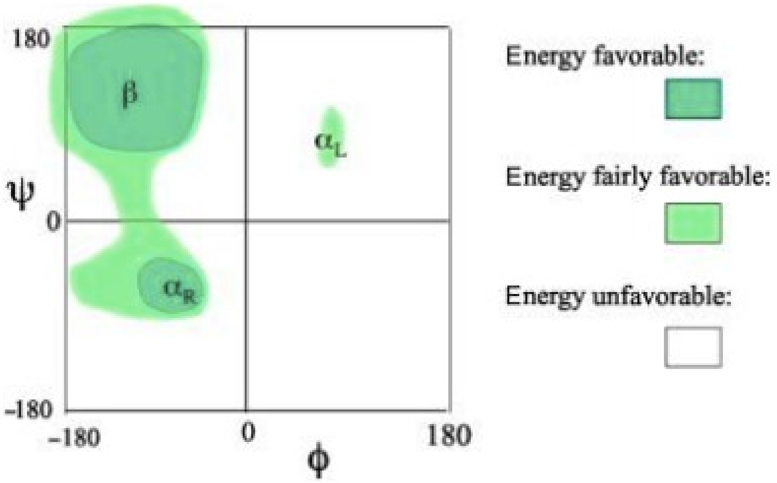
3.4 Validity assessment of structures of unknown origin

- Check parameters for which we know how they should look like.
- Methods to determine what is normal:
  - Data driven approach: examine known structures, do statistics
  - Theoretical approach: perform calculatioons:
    - ab initio: based on quantum mechanics
    - Molecular mechanics
- If nothing fits, the structure is crap. If everything fits, probably over refined. If 5-10% don't fit, it can be OK, there are always parts that look different.

3.4.1 Ramachandran map

Not all combinations of two characteristic torsion angles of the polypeptide chain (backbone atoms) are favorable.

- One of the most sensitive structure validation tools
- D-amino acids have opposite conformational preferences



3.5 Other parameters in structure validation

- Atomic contacts: both density and specific contacts should match what is common: no close contacts should occur
- Secondary structure: to few helices and strands in globular proteins might indicate problems
- Hydrogen bonds: donors and acceptors should be paired in protein interior, should match
- side-chain torsion angle preferences
- Deviations from planarity (amide planes, aromatic rings, etc.)
- Typically no cavities inside structure unless functional
- For structures containing only alpha carbons: interdependence of angles and torsions

4 Secondary structure assignment

4.1 Protein structure hierarchy

- **Primary structure**: sequence
- **Secondary structure**: turns, loops, α-helix, β-strands, sheets. The local conformation of the polypeptide chain
- **Tertiary structure**: 3D fold
- **Quaternary structure**: subunits

Secondary structure is the local structure of the polypeptide chain. The term secondary structure is often restricted to α-helices and β-sheets. Related structures often exhibit similar secondary structure.

The importance of secondary structure lies in

- Structure visualization
- Characteristic of the fold
- May affect sequence alignment
- Has functional consequences
- Key concept in comparative structure analysis
- Secondary structure prediction is an important part of teritary prediction
- The accuracy of second structure prediction from sequence is so high that is meets that of its assignment from 3D structure.
- Description of folds: list of relative orientation of secondary structure elements is key
- Comparing related folds can be easily visualized using topology diagrams.

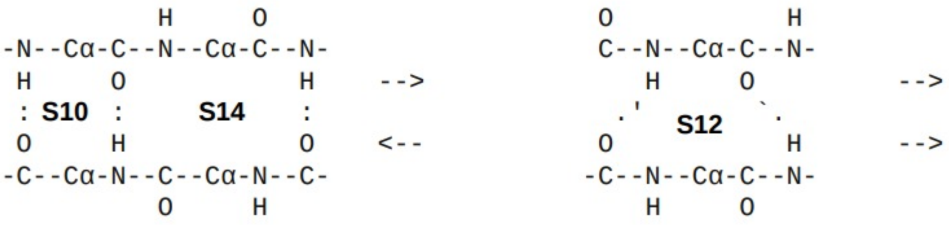
4.2 Special types of secondary structures

- The coiled coil: a supersecondary structural element
  - Specific interactions between α-helices: packing of side chains
  - Underlying amino acid sequence: repeating units
  - Not to be confused with other α-helical structures
- The polyproline II helix:
  - Characteristic of collagen
  - Tripple helix

4.3 Assigning secondary structure

- Pauling's original definition is based on H-bonding pattern.
  - Helices:
    - α-helix: i–$i+4$ H-bonding: 3.6 residues per turn, 13 atoms in H-bonded pseudoring. 1.5A rise per residue.
    - 3_10 helix: i–$i+3$
    - Π-helix: i–$i+5$.
  - Strands and sheets:
    - Strand: one continuous chain segment
    - Sheets: segments held together by H-bonds
      - Parallel: 12 atoms in H-bonded pseudoring

· Antiparallel: 10/14 atoms in H-bonded pseudoring. This is more favourable energetically



- Regions of the Ramachandran map
  - Hydrogen bonding can be formed only by residues adopting particular backbone conformations
  - H-bonding patterns and backbone torsions are two manifestations of secondary structure.
  - Residues in the given Ramachandran region might not form the characteristic hydrogen bonds, it dosen't matter how we assign secondary structure elements.

We can categorize conformations based on backbone torsion angles. Atypical structures don't have defined secondary structures, while typical structures can be further divided into two groups: periodic or homoconformers (helixes, β strand, type III β-turns) and aperiodic or heteroconformers (I, II, VI, VIII, γ turns).

4.4 Assigning hydrogen bonds

Amide H coordinates often missing from X-ray structures, yet naturally we need them to analyze hydrogen bonding.

4.4.1 Criteria

Distance-angle criteria

$$\theta > 120°, \quad r_{HO} < 2.5A,$$

where θ is the N-H-O angle and r_{HO} is the H-O distance.

Coulomb energy criterion (DSSP)

$$E = f\partial^+\partial^- \left(\frac{1}{r_{NO}} + \frac{1}{r_{HC'}} - \frac{1}{r_{HO}} - \frac{1}{r_{NC'}}\right),$$

where $f = 332A\frac{kcal}{e^2 mol}$, ∂^-, ∂^+ are $0.2e, -0.42e$ respectively. Energy cutoff is $-0.5kcal/mol$. Does not consider atom-atom repulsion and does not result in a characteristic H-bond length.

Empirical criterion (STRIDE)

$$E_{hb} = E_r + E_t + E_p,$$

where E_{hb} is the total enery of the H-bond, E_r is the N-O distance, E_t and E_p depend on the relative positions of the O, H, N atoms and the N-C'-O amide plane.

4.5 Assigning secondary structure with programs

4.5.1 DSSP – Definition/Dictionary of Secondary Structure Proteins

- de facto standard in secondary structure assignment
- H-bonds only
- Detailed output
- Secondary structure assigned to each residue:
  - H: α-helix

  - G: 3_10-helix
  - I: Π-helix
  - E: β-strand
  - B: β-bridge
  - T: turn
  - S: bend

- α-helices: at least two consecutive i-$i+4$ bonds required at its beginning and should end with two consecutive $i-4$-i H-bonds.
- Similar criteria for I and G.
- DSSP will not classify the first and the last residues involved in H-bonding.
- Segments with one such H-bond are classified as T.
- β-strands: residues should be involved in either two appropriate H-bonds or flanked by two such bonds
- Minimal β-strand is at least two residues per strand.
- S is defined by geometric criteria

4.5.2 STRIDE – secondary STRuctural IDEntification method

- Considers backbone torsion besides H-bonds
- Assigns helical and sheet propensity to backbone conformations based on their closeness to ideality
- STRIDE was optimized to conform to a set prepared by experts
- Structural assignment by STRIDE is similar to that of DSSP
- STRIDE might include terminal residues in the absence of H-bonding when the torsion angles are OK
- Identifies turns based on torsion angles, assigns C to them

4.5.3 Practical aspects

Secondary structure assignment depends on the method used. First and last residues can be ambiguous. Comparison with secondary structure predicted from sequence is not straightforward.

4.5.4 Considering DSSPcont

Secondary structures of the same or closely related proteins differ in structure from different experiments. This can be caused by

- Different condition (T, pH, etc)
- Experimental error
- Thermal fluctuations

DSSPcont aims to minimize the effect of these. The authors claim secondary structure calculated for a single NMR conformet by DSSPcont captures the variability of DSSP results on all conformers.

5 Assignment of structural domains

There are proteins consisting of functionally and or structurally distinct parts. Sometimes these are evolutionary units also. In general, a domain is a region of a protein with specific properties. These properties can be structural, functional or both. There are a number of different approaches to find and characterize domains depending on what exactly we look for. Domains may or may not have globular structure.

The term motif covers a similar concept typically used for shorter, recurring segments.

Examples include

- Proteins responsible for cell-cell interactions in multicellular animals
- Domains can be responsible for interactions with specific partner molecules. These can be peptides, or other domains.

5.1 Characterising domains

5.1.1 Experimental approach

- **Local approach**: dissecting the protein to regions
- **Global approach**: investigating the role of the region/domain in the context of the full protein

5.1.2 Computational approaches

- **Sequence based**: Key concept: similarity
- **Structure based**: Key concept: compactness - domains should be seperated from other parts of the protein. Only applicable to globular domains. Intra-domain contacts are more intensive than inter-domain contacts.

Functional/structural similarity, with complex enough sequences implying evolutionary relationships.

5.2 Domains in structural bioinformatics

In structural bioinformatics, globular domains constitute the basic units of operations like

- structural alignment
- structure comparison and classification
- functional assignment
- 3D structure prediction

We expect that domains will be biologically meaningful, and to make sense in terms of structure (being compact, contain uninterrupted regular secondary structure elements in contact with each other). Domain definitions are consistent with each other: different structure and sequence based methods yield similar domain assignment.

There are problems with domains.

- Domains might not be units neither in terms of function or autonomous folding.
- Structural units might not correspond to a continuous segment of the polypeptide chain.

5.3 Domain assignment methods

5.3.1 First–generation methods

These methods are not capable of global domain partitioning. This results in a hierarchic domain assignment: split the chain into two then split the domains further if applicable.

- Hydrophobic cores and folding clusters
- interface area for different cuts

5.3.2 Second generation methods

Capable of global partitioning irrespective of covalent structure.

- **STRUDL**: Based on contact area between groups of residues.
- **DomainParser**: Graph-based method, applying the Ford-Fulkerson algorithm.

  Represent each residue as a node in the graph. Represent the contacts between residues as edges connecting nodes: strength of the interaction between the two residues is reflected by the capacity of the edge connecting two nodes.

6 Structure comparison and alignment

Structure is more conserved than sequence. Related proteins differ in sequence but have the same fold. Most mutations do not affect global structure. Proteins with similar structure have some key residues conserved even when the rest of the sequence is not. These are in fact evolutionary related but this can not be established by sequence comparison alone.

Structural rearrangements can occur when proteins are performing their function. Identifying and quantifying these requires structure comparison.

6.1 Important definitions

Structure superposition Finding the best match between two sets of corresponding 3D points. Has an exact mathematical solution (e.g. least squares fitting).

Structure alignment Find the correspondence between two sets of 3D points that give the best fit. This is an NP-hard problem. Genuine structure alignment ignores the sequence.

Structure comparison Describe the extent of similarity between two structures. May or may not include structure alignment.

All of these can be local or global. The best mathematical solution might not be the biologically most relevant. As with sequences, we can use structure comparison methods to search databases for similar structure and can also perform multiple alignment.

Similar concepts can be demonstrated on sequences. In sequence superposition for sequences differing in only one position, we can simply write them below each other.

In sequence alignment, with sequences with no trivial similarity, we need to search for the best correspondence of the positions.

In sequence comparison, we can utilize general features like amino acid composition etc. Alignment might not be feasible here but the two segments share similar characteristics.

6.2 Structure superposition

In the simplest case, the structures are identical and the correspondence between atoms is trivial, for example with NMR ensembles (different conformers of the same molecule).

Usually a measure of goodness here is the root-mean-square deviation (RMSD) of the atomic positions. The mathematical algorithms we use to get a good result include the Least-squares fit algorithm, and ML-optimized fit.

6.3 Structure alignment

In structure alignment we have to determine the point-to-point correspondence between the two structures. Usually three key steps are performed:

1. **Representation**: How to represent the input structures in a coordinate-independent manner suitable for alignment

2. **Optimization (search)**: How to sample the space of possible alignment solutions between the structures

3. **Scoring**: How to score a given alignment and determine its statistical significance (e.g. Z-score)

6.3.1 Methods

Structure alignment is NP-hard. There are many algorithms, that differ in the three key steps; representation, search and scoring, as well as the data set used for development, models available, used database and implementation.

Overview of structural representations and methods for selected servers

- **Dali**:
  - Representation: distance matrix (C_α distances)
  - Search method: find overlap between (sub) matrices
- SSAP:
  - Representation: C_β-C_β vectors
  - Search method: double dynamic programming
- VAST
  - Representation: secondary structure elements
  - Search method: graph theoretical approach
- CE
  - Representation: 8-residue segments
  - Search method: identify matching segments, extend matching regions
- SALIGN
  - Representation: User-defined arbitrary descriptors
  - Search method: dissimilarity matrix, dinamyc programming
- FATCAT
  - Representation: C_α atoms of protein fragments
  - Search method: chaining of aligned fragments using dynamic programming

6.4 Multiple structure alignment

In multiple alignment we start with pairwise alignments. These pairwise alignments can be added

- iteratively, by adding novel structures
- tree-based approach
- Monte Carlo optimization of the full multiple alignment

As with sequence alignments, multiple structure alignments are more informative in terms of conserved, variable sites, regions.

Usually only the core regular secondary structure elements (helices, sheets) are aligned in a large family.

7 Uses of structure comparison: classification and function assignment

Applications of structure comparison include fold classification and prediction of function.

The term **homology** means evolutionary relationship. Two proteins, genes or organs, etc. are homologous if they can be traced back to a common ancestor Thus, it does not have a level or degree, although you can often read sentences like 'these two proteins share 60% homology'. This is correctly formulated as 'these two proteins are homologous and share 60% sequence identity or similarity'.

Homology is a purely evolutionary term and does not necessarily imply any functional or other similarity (**analogy**), just refers to the history of the objects. Function can be gained and lost during evolution, and similar function does not imply homology.

For example the wings of a bird and a bat are homologous as a whole (evolved from the forelimb of their last common ancestor), and they are also analogous as a whole, as the wings are used for flight. However different parts of the wing are not necessarily homologs (eg. wing tips) and vice versa: homologous bones may be responsible for different function, meaning the two wings evolved differently (convergent evolution).

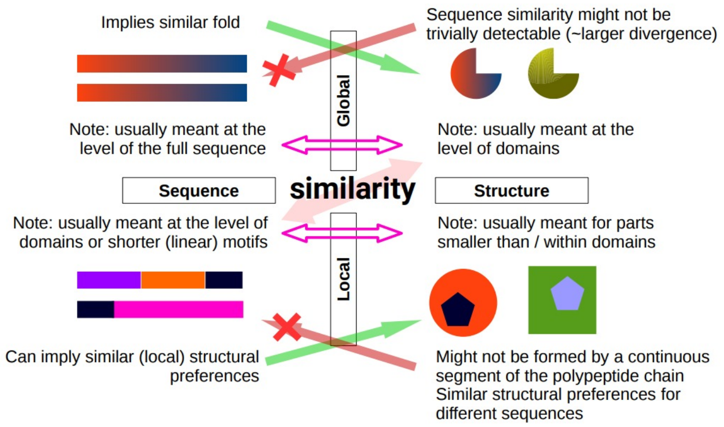
It's the same with proteins. Evolutionarily related proteins might have different functions, and different global fold with similar local structure implies different evolutionary history but similar function: convergent evolution.

7.1 Cases of homology: orthology and paralogy

Orthologous genes/proteins basically reflect the history of species, whereas paralogs arise by gene duplication.

For example, all globins are homologs. All myoglobins are orhologs of each other, and all hemoglobins are also orthologs to each other. Hemoglobin an myoglobin are paralogs and hemoglobin itselt contains 2 paralogous subunits (α and β).

7.2 A general strategy to extract function from similarities


Implies similar fold
Sequence similarity might not be trivially detectable (~larger divergence)
Global
Note: usually meant at the level of the full sequence
Note: usually meant at the level of domains
Sequence
similarity
Structure
Note: usually meant at the level of domains or shorter (linear) motifs
Note: usually meant for parts smaller than / within domains
Local
Can imply similar (local) structural preferences
Might not be formed by a continuous segment of the polypeptide chain
Similar structural preferences for different sequences

We can use global similarity to detect evolutionary relationships: identification of possible orthologs and paralogs. This can provide a first guess on function.

Then we use local similarity to check the presence of functional sites. Thus, the presence and the absence of sites can justify or falsify guesses on global similarity and/or provide clues undetectable from evolutionary relationships.

7.3 Structure classification

For structure classification, the following data is needed:

- Structural data as input
- A unit to classify: domain
- Methods to compare structures (previous section)

- A classification scheme (eg. hierarchic)
- A purpose to be fulfilled with the classification

The problem with structure classification include:

- Structure comparison algorithms are generally slow
- The number of available structures grows exponentially
- It is not realistic to recalculate everything for newly released structures
- It is much faster to assign proteins to classes based on their sequences

7.4 Major structure classification databases

7.4.1 CATH

This database is named after the first 4 levels of classification: Class, Architecture, Topology, Homology. Homology is used most often to define relatedness.

CATH includes a structure viewer where the domain can be viewed in context. It provides a structural classification along with functional information on class members.

Level H is considered to be evolutionarily relevant, proteins at the same H level are homologs. Higher level groups do not necessarily contain structures that are all related to each other.

7.4.2 SCOP

The acronym of Structural Classification Of Proteins. The levels of hierarchy in this database are: Class, fold, superfamily and family. Proteins in the same family are thought to be evolutionarily related. Some superfamilies might also reflect evolutionary relations.

SCOP2 allows seperate handling of evolutionary relationships and structural similarities. Proteins are represented on a directed acyclic graph. The database includes globular, transmembrane, fibrous and disordered protein types.

7.5 Protein function assignment

The meaing of function depends on the level of understanding: What impariment of the gene can cause? What pathway is the protein involved in? What is its enzymatical function? etc...

Also proteins may have multiple functions. For example phosphoglucose isomerase in its intracellular for is dimeric and is involved in glycosis, andin extracellular form it is monomeric and is involved in signal transduction.

Gene ontology (GO) is a hierarchic controlled vocabulary to denote cellular localization, biological processes and molecular function.

The enzyme catalog (EC) assigns a four number code to each enzyme. It classifies reactions, not proteins, meaning that unrelated proteins can catalyze the same reaction. It contains no information about the mechanism of the reaction, and is uninformative in many aspects. This is still widely used, as reactions can be classified much better than other functions.

7.6 Local structural similarity

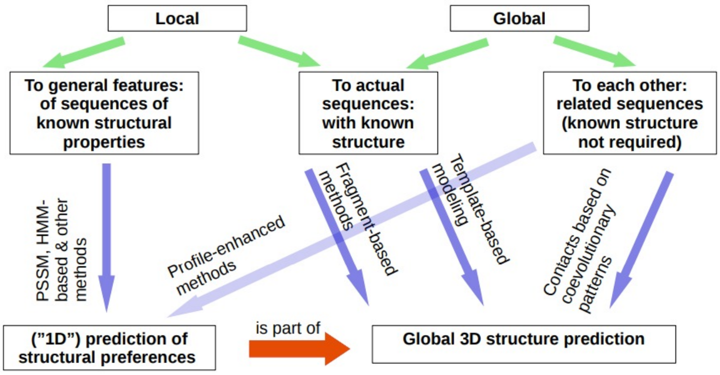
We compare known functional 3d motifs to unknown ones based on spatial/physiochemical properties.

- SiteEngine: comparison ATP-binding sites
- ProBis: Heme-binding sites
- CMASA: catalytic triad of serine proteases

8 Structure prediction

Structural information is more informative than sequence alone. However experimental structure determination is a lot of work and a lot of money, and can't be guaranteed to be successful in many cases. It is especially hard for special cases - most easier ones are already solved.

Prediction is cheap and fast, but relies on similarities and/or physical considerations. Recognition of similarities is not always trivial, our predictions will only be as good as our understanding of the similarity. It is easy to obtain some output, but how do we know whether the results are valid?


Local
Global
To general features: of sequences of known structural properties
To actual sequences: with known structure
To each other: related sequences (known structure not required)
PSSM, HMM-based & other methods
Profile-enhanced methods
Fragment-based methods
Template-based modeling
Contacts based on coevolutionary patterns
("1D") prediction of structural preferences
is part of
Global 3D structure prediction

8.1 Structure prediction in 1D

Predictions in 1D assign a state/value to each residue.

- Based on some kind of local similarity
- Secondary structure (Helix, Extended sheet, Coil)
- Specific structural motifs (eg. coiled coils)
- Intrinsic disorder (disorder tendency score)
- Transmembrane helices/topology (tm helix/in/out)
- Solvent exposure
- Aggregation propensity

Residue distribution in disordered segments can be a good basis for structure prediction. Characteristic differences compared to globular proteins include high fraction of charged residues, Gly and Pro, and depleated hydrophobic, especially aromatic residues and Cys.

Tendencies of amino acid residues to be located in specific regions of different structural preferences/properties (enriched/depleted):

- disorder (IDS±)
- transmembrane helix forming (TMH±)
- aggregation propensity (AGR±)

8.1.1 Principles of predicting secondary structure

The typical output is helix (H), sheet (E), other/coil (C). The Chou-Fasman method, which is the earliest method, they take into consideration the residue level, and abundance of amino acids in sheets and helices.

Modern methods incorporate evolutionary information, they identify similar sequences, generate alignment. For this they use weighted alignments and neural networks. This yields results optimized for the protein family investigated. Most of the new methods are accurate at the level where secondary structure assignment from structure becomes ambiguous.

8.1.2 Position–specific scoring matrices (PSSMs)

The principle of the method is to take an alignment of segments known to have a specific property and generate an abundance matrix of the alignment. We then transform the matrix ito a weight or some score matrix that can be used to score any given sequence aligned to it. The score can be used to assess whether a given sequence / segment belongs to the group with the given property or not.

8.1.3 Hidden markov models (HMMs)

The principle of this method is to take an alignment of segments known to have a specific propery and generate a probabilistic model with multiple internal states describing the

- probabilities of transitions between states
- probabilities of emitting given signals in given states
- signals can be amino acids or structural properties

We can use this model to assign a probability to a query sequence. This probability will tell us whether the query sequence belongs to the group or not.

8.1.4 Specialized predictions

IUPred for intrinsically disordered segments The principle is to estimate the interaction energy of residues within a segment. Uses a simplified energy function to calculate the energy in globular proteins. A predictor was built that estimates these energies from sequence. We can use this algorithm to detect segments with high interaction energies.

IUPred2A for intrinsically disordered segments This method offers integrated prediction and recognition of

- Intrinsic disorder
- Presence of disordered binding sites based on estimated enery in the putative bound state
- Pfam domains
- Linear motifs based on the ELM (eukaryotic linear motif) database
- Post-translational modifications
- Disordered regions annotated by experimental data
- Available PDB structures corresponding to regions

Machine learning in the prediction of structural features Deep neural networks are increasingly used for various structure prediction tasks like disorder and secondary structure. Such methods work well if there are plenty of data available both in terms of different inputs and features in each input. They usually use several residue features as input. They generally make use of information on evolutionarily related sequences in the form of PSI-BLAST (Position-specific iterative basic local alignment search tool) profiles and HMMs. The more homologs, the better.

PSI-BLAST derivers a position specific scoring matrix (PSSM) or other profile from the multiple sequence alignment of sequences detected above a given score threshold using protein-protein BLAST. This PSSM is used to further search the database for new matches, and is updated for subsequent iterations with these newly detected sequences. THus, PSI-BLAST provides a means of detecting distant relationships between proteins.

8.1.5 Prediction evaluation

Secondary structure Different methods will yield different results. It is always a good practice to look for consensus. Different methods also have different reliability. The fact that a method performs well on most proteins does not mean that it cannot give false results for special cases.

Sometimes we get contradictory results from different methods. We have to use common sense and biological knowledge to judge.

Disordered segments Performance of different methods on annotated, experimentally validated disordered segments listed in the DisProt database.

The four options are

- TP: true positives (redisues located in disordered segments and predicted as such)
- TN: true negatives
- FP
- FN

We define two properties.

$$\text{sensitivity} = \frac{TN}{TP + FN},$$

which defines how well do we find the correct segments, and

$$\text{specificity} = \frac{TN}{FP + TN},$$

which defines how well we found only the correct segments.

There are several motifs, that are regularly predicted as disordered but they are in fact structured. In such cases more specialized predictors have precedence over more general ones when evaluating the results.

Examples of structural motifs that are often predicted to be intrinsically disordered:

- Coiled coils
- Single $\alpha helices$
- Collagen triple helices

8.2 Structure prediction in 3D

Predictions with 3D information use dissulfide connectivity and beta-barrel transmembrane proteins. Full 3d predictions have multiple methods depending on the level of similarity to proteins with known structures.

8.2.1 Evaluating models

TM-score The TM-score (template model) is one of the most commonly used scores today

$$\text{TM-score} = \max \left[\frac{1}{L_n} \sum_{i=1}^{L_T} \frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} \right],$$

where L_N is the length of native structure (template), L_T is the number of residues aligned to the template, d_i is the distance between the ith residue pari, d_0 is a normalization factor and max indicates that the best score over a series of possible alignments.

The aim in its development was to eliminate dependence from length, radius of gyration, etc.

Needs a structure alignment. In practice, this is performed iteratively to obtain an optimal one. Not all parts of the structure will be modelled equally well. Models with a TM-score of over 0.5 are generally regarded as correct.

GDT-TS (Global distance test - total score) In GDT the algorithm identifies in the prediction the sets of residues deviating from the target by not more than specified CA DISTANCE cutoff using many different superpositions. Each residue in a prediction is assigned to the largest set of residues deviating ftom the target by no more than a specified distance cutoff. This measure can be used to evaluate ab-initio 3D and comparative modelling predictions.

$$GD - TS = (GDT - P1 + GDT - P2 + GDT - P4 + GDt - P8)/4,$$

where $GDT - Pn$ denotes the percent of residues under distance cutoff $\leq A$.

8.2.2 3D structure prediction from template

Applicable when a protein with known structure shows cleas sequential similarity to target. A high degree of structural similarity is expected. Distinct approaches

- Building target based on overlapping fragments of known structures, then build different parts.

  Loop modelling: have to build anew or find a known segment that matches the length of the target loop, is similar in sequence, bridges a 3D distance appropriate for the target structure and is physically feasible.

- Optimize interatomic distances based on the templates

SWISS-MODEL is freely and easily accessible and now can also build models of protein complexes. For this, the method uses the structure of homologous complexes and features of binding interface conservation.

8.2.3 3D structure prediction from scratch

The ideal case is that we would be able to build models based on physical principles alone. This is not feasible with quantum chemistry. Forcefield based methods are not nearly this accurate and the computational time for exhaustive conformational sampling is still a strong limitation.

To get an accurate model, we use an appropriate combination of

- **Threading**: Identify known structures compatible with sequence to be modeled. In principle we test wether a known fold can be adopted by the target sequence.
- **Fragment-based modelling**: Find fragments with known structure matching the sequence of the target, then join these fragments, and model unmaped parts. We then optimize the full modeled structures. This can be done with hybrid approaches, which use experimental data, which is traditionally not sufficient for high quality structure determination.
- **Modelling based on predicted contacts**: Get as many related sequences as possible. Identify coevolving residue pairs, filter those likely to form direct contacts and transform contact information to restrain or select structural models obtained with a suitable sampling method.

I-TASSER template detection I-TASSER searches a PSI-BLAST profile generated for the input sequence against precomputed PSI-BLAST profiles.

8.2.4 Predicting structures based on evolutionary information

This methid still uses sequence similarity in spite of no similarity to any known structure that is used. The more known sequences, the better. To get meaningful results the sequence alignment should reflect the equivalent positions in the structure.

Possible indirect contacts are filtered out with an elaborate statistical approach. The resulting distance constraints are ranked, more likely ones with higher precedence. Cysteines are allowed to pair with other cys.

Secondary structure prediction is part of the procedure. Correlation-inferred distance restraints between residues in the same predicted elements are eliminated from subsequent steps.

9 Protein–ligand docking

Modelling a protein-ligand interaction is actually an optimization, as we assume, that the bound structure corresponds to an energetic minimum (docking). Aims of a docking study may include

- Determine whether a given ligand can bind to the receptor
- Determine which ligand binds the receptor best
- How selective a given receptor-ligand interaction is
- Determine the structure of multisubunit protein complexes

Docking can be useful to filter out ligands to be examined experimentally nad in the design of better ligands based on the modelled interaction.

9.1 General overview of molecular docking

One of the partners is a biomacromolecule, the othes can be a small molecule or another macromolecule. Docking can be combined with experimental data.

Applications of small molecule docking applications:

- **Target fishing and profiling**: Prediction of targets for compounds on the basis of ligand-receptor complementary.
- **Prediction of adverse drug reactions**: Prediction and rationalization of drug off-target activities on the complementary between ligands and targets
- **Polypharmocology**: Identification and optimization of compounds that simultaneously modulate a set of targets involved in the same disease.
- **Drug repositioning**: Identification of novel therapeutic-relevant targets for already marketed drugs, and known chemicals and natrual entities
- **Ligand-target binding rationalization** Identification of the structural determinants necessary for the efficient ligand-receptor binding
- **Virtual screening** Identification of compounds modulating disease-related targets and their optimization

In signaling protein like a transmembrane receptor, ligand-induced conformational changes are required for signal transduction, these are not induced by antagonists. In terms of chemistry, there are many kinds of ligands.

In the context of docking, the protein with the binding site is referred to as receptor. The ligand can also be a protein.

The ligand binding site on the receptor can be known or unknown. For example in blind docking the determination of the binding site is part of the task.

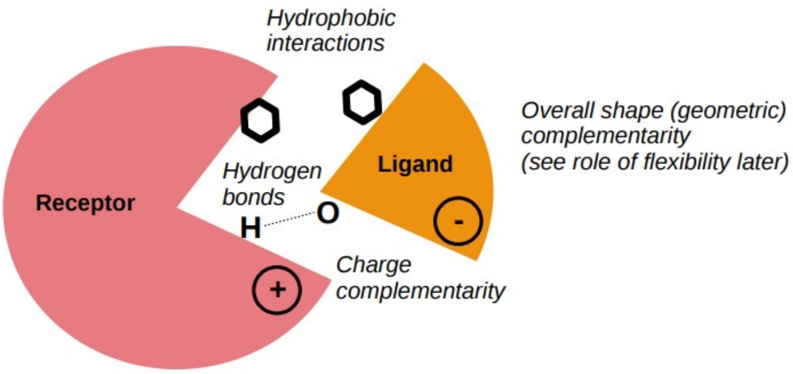
The structure of the receptor can be known or unknown. The two types of docking methods:

- **Single receptor-ligand pairs**: Can use more details and can evaluate multiple solutions.
- **Screening of large ligand sets**: Should be optimized for speed, and yield simple measures.

The output of a docking is a set of poses (protein-ligand interaction geometries) with corresponding scores and or energies.

Protein-ligand complementary is the basic concept in docking, this is the target function to be optimized during the calculations.

Not only geometry, but also the physiochemical properties of the properly positioned ligand should match those of the receptors.



The problem is that we do not necessarily know which of the surface features of the molecules should match the best, or are biologically relevant.

9.2 General steps in docking

1. Create a suitable representation for the receptor and the ligand:
   - Ligand: Can be handled explicitly (all atom representation) and the flexibility can be taken into account (rotatable bonds)
   - Receptor: simplified representation needed, with focus on the binding site. Only limited flexibility can be modeled explicitly. Grid-based representations can efficiently capture many aspects.
2. Search the space of possible solutions, while being efficient yet exhaustive.
3. Evaluate and rakn solutions

9.2.1 Search methods

The aim is to obtain a number of relevant poses to be scored. To do this we have to define what to search for.

We search for ligan conformations and mutual orientation of receptor and ligand.

To do the search, the receptor is fixed, while the ligand is being moved. To recognize interactions we create interaction spheres around sites and match ligand sites to grid positions.

In this search we have to consider ligand flexibility. We need to generate conformers, and try them all. We can try to optimize the ligand geometry to the binding site, or build up the ligand within the site, fragment by fragment.

9.2.2 Principles of scoring

Our aim when scoring is to effectively discriminate between relevant and irrelevan poses. For this, there are three common methods:

- **Force field-based methods** THis is based on a set of equations that describe molecular conformation using an approximation with principles of classical physics.
- **Empirical methods**: Intuitively select interaction terms (H-bonds, entropy, etc.), create equations and optimize parameters on a test set.
- **Knowledge based**: The assumption here is that frequently observed contacts must be favorable.

Not only the score but also the similarity of the poses can be informative. Finding the pose with the highest biological relevance, etc.

Water molecules might form hygrogen bonds with the ligand and the receptor, providing means of indirect interaction. This is very hard to predict. If the experimental structure of the receptor contains water molecules in the binding site , we might leave them there, otherwise we might add water molecules to suitable positions within the site and then try ligand poses.

It is not uncommon to use modeled receptor structures, but it needs extra care, as the details of the modeled binding site are usually among the least reliably modeled aspects of a structure. If possible, it needs to bee validated by other computational methods and, if possible, experimental data.

9.2.3 Flexibility of the receptor

Proteins are dynamic. The ligand-bound state might differ from the free one. Complexes with different ligands might also be different. The extent and details of structural change is very hard to predict.

This can be done a couple of ways. In soft docking we allow overlap, and try to optimize the receptor geometry during docking. We use an ensemble based representation for the receptor and examine its structure with different ligands, including similar ones to the one being docked.

9.2.4 Special docking applications

Fragment based ligand discovery is based on the idea to dock small fragments and see whether they can be reasonably linked to form a larger molecule as an effective ligang. There are experimental methods with a similar strategy.

Covalent docking is when the ligand forms a covalent bond with the receptor. Special docking methods needed to take this into account as this means a special, constrained interaction.

Exhaustive multisite docking is to predict and analyze multiple binding sites to analyze possible allosteric phenomena. 'Wrap'n'shake' is multiple rounds of blind docking untill the whole surface is covered, then remove the weak blinders.

10 Structural ensembles to describe protein structure and dynamics