

KEMM15

Lecture note in structural bioinformatics: A practical guide

S Al-Karadaghi, Biochemistry & Structural Biology, Lund University

BASICS OF PROTEIN STRUCTURE	3
SOME DEFINITIONS	3
THE 20 AMINO ACIDS AND THEIR ROLE IN PROTEIN STRUCTURES AND FUNCTION	4
TORSION ANGLES AND THE RAMACHANDRAN PLOT	6
THE RAMACHANDRAN PLOT AND THE QUALITY OF A PROTEIN STRUCTURE	8
SECONDARY STRUCTURE ELEMENTS	8
STRUCTURAL MOTIFS: CONNECTIVITY BETWEEN SECONDARY STRUCTURE ELEMENTS	11
FOLDS AND FOLD CLASSIFICATION	13
DOMAINS AND DOMAIN CLASSIFICATION	15
PROTEIN DATABASES: SHORT OVERVIEW	17
THE PROTEIN DATABANK (PDB): FILE FORMAT AND CONTENT	21
SEQUENCE ALIGNMENT AND ANALYSIS	24
OVERVIEW	24
SEQUENCE ALIGNMENT BASICS	24
AMINO ACID SUBSTITUTIONS AND AMINO ACID REPLACEMENT MATRICES (PAM, PET91, BLOSUM)	26
SEQUENCE ALIGNMENT TUTORIAL 1	28
SEQUENCE ALIGNMENT TUTORIAL 2: BCHI-BCHD ALIGNMENT	31
INTRODUCTION TO HOMOLOGY MODELING	34
OVERVIEW	34
PROTEIN HOMOLOGY MODELING USING THE SWISS MODEL SERVER	35
STEP BY STEP MODELING	36
QUALITY ASSESSMENT OF A HOMOLOGY MODEL	38

Basics of protein structure

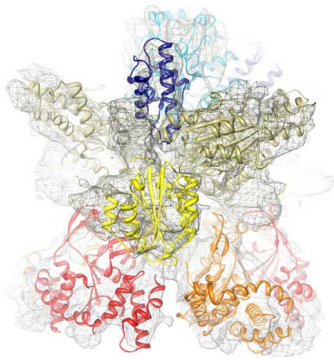
Some definitions

To understand the basic principles of protein three-dimensional structure and the potential of their use in various areas of research, academic or industrial - like pharmaceutical or biotech industries - we first need to look at the four levels of protein structure. The different structural levels depend on each other, together creating an extremely complex network of interactions between hundreds and thousands of atoms. The **first level** is the amino acid sequence - there are 20 different amino acids most commonly found in proteins. The amino acids are joined to each other into a polypeptide chain during the process of protein synthesis essentially. The sequence controls to a large extent the higher levels of the protein structure – **secondary**, **tertiary** and **quaternary** structure. The tertiary structure is essentially the way by which secondary structure elements are arranged in space in different structural **motifs**, **folds** and **domains**.

A **domain** is an independent folding unit of a protein. It is independent because domains may often be cloned, expressed and purified independently of the rest of the protein, and they would still show some characteristic activity, like ligand binding, metal binding or interaction with other proteins or even with other domains of the same protein. Some proteins consist of one single domain while others may contain several domains. A protein domain is assigned a certain type of fold. Domains with the same fold may or may not be related to each other functionally or evolutionary. This is because Nature appears to have re-used some protein folds multiple times in different contexts. The currently known protein three-dimensional structures have been classified into more than 1000 different unique folds. In the following chapters we will discuss some examples of these folds, to illustrate the basic principles used for their definition.

The fourth structural level, the **quaternary structure**, is an oligomeric structure and usually involves several polypeptide chains (called **subunits**). It may be the same protein molecule (**homo-oligomer**) or different protein molecules (**hetero-oligomer**). An oligomer is stabilized by subunit interactions, and may involve hydrophobic interactions, hydrogen bonds, salt bridges, etc. The different molecules within an oligomeric structure may contribute to an active site (or sites), contribute to the dynamics of the complex and may interact with some target proteins outside the complex.

Since large variations in the sequence may result in the same type of three-dimensional structure, we say that **structure has a higher degree of conservation than sequence**. This can be understood if we take into account function – for example binding of a certain ligand, specificity of interactions with other proteins, dynamic behavior of a structure – all depend on the type of the structure. This is why you may hear that the determination of the structure of a protein with unknown function may help in revealing the function. An interesting example was provided by the anaerobic cobalt chelatase, an enzyme active in vitamin B12 synthesis. Although the function of the protein was known before structure determination (Schubert et al., 1999), the similarity of the structure to that of ferrochelatase (Al-Karadaghi et al., 1997), an enzyme active in heme biosynthesis, could only be revealed after the structure determination of cobalt chelatase. The reason is that there is only 11% sequence identity between the two proteins, a number much smaller than the so-called "homology-threshold", normally considered in sequence alignment to be an indication of the existence of evolutionary relationships between proteins (around 20-25%, will be discussed in a later chapter).



An example of a quaternary protein structure. The figure shows the complex of two of the subunits of the enzyme magnesium chelatase. The structure was obtained using single-particle reconstruction from cryo-electron microscopic (cryo EM) images of the complex. Where appropriate, the available X-ray structure of subunit BchI of the enzyme was docked into the EM density (shown in ribbon representation). Other domains were homology-modeled based on known structures from other proteins. Published in Lundqvist et al, Structure 2010.

The 20 amino acids and their role in protein structures and function

The amino acids are put together into a polypeptide chain on the ribosome during protein synthesis. In this process the peptide bond, the covalent bond between two **amino acid residues**, is formed. There are **20 different amino acids** most commonly occurring in nature. Each of them has its specific characteristics defined by the side chain, which provides it with its unique role in a protein structure. Based on the propensity of the side chain to be in contact with polar solvent like water, it may be classified as hydrophobic (low propensity to be in contact with water), polar or charged (energetically favorable contact with water). The charged amino acid residues include lysine (+), arginine (+), aspartate (-) and glutamate (-). Polar amino acids include serine, threonine, asparagine, glutamine, histidine and tyrosine. The hydrophobic amino acids include alanine, valine, leucine, isoleucine, proline, phenylalanine, tryptophane, cysteine and methionine. You probably noticed that this classification is based on the type of the amino acid side chain. However, glycine, being one of the common amino acids, does not have a side chain and for this reason it is not straightforward to assign it to one of the above classes. Generally, glycine is often found at the surface of proteins, within loop- or coil (without secondary structure) regions, providing high flexibility to the polypeptide chain at these locations. This suggests that it is rather hydrophilic. Proline, on the other hand, is generally non-polar and is mostly found buried inside the protein, although similarly to glycine, it is often found in loop regions. In contrast to glycine, proline provides rigidity to the polypeptide chain by imposing certain torsion angles on the segment of the structure. The reason for this is discussed in the section on torsion angles. Glycine and proline are often highly conserved within a protein family since they are essential for the conservation of a particular protein fold.

Below the 20 most common amino acids in proteins are listed with their three-letter and one-letter codes:

Charged (create salt bridges, charge-charge interactions):

- Arginine - Arg - R
- Lysine - Lys - K
- Aspartic acid - Asp - D
- Glutamic acid - Glu - E

Polar (may participate in hydrogen bonds):

- Glutamine - Gln - Q
- Asparagine - Asn - N
- Histidine - His - H
- Serine - Ser - S
- Threonine - Thr - T
- Tyrosine - Tyr - Y
- Cysteine - Cys - C
- Tryptophan - Trp - W

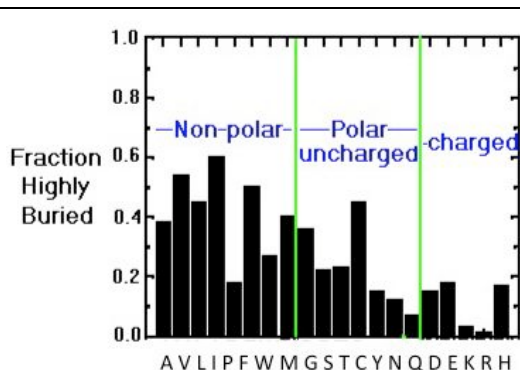
Hydrophobic (normally buried inside the protein core):

- Alanine - Ala - A
- Glycine - Gly - G
- Isoleucine - Ile - I
- Leucine - Leu - L
- Methionine - Met - M (although it may accept hydrogen bonds in some cases, see <http://www.biochem.ucl.ac.uk/bsm/atlas/met.html>)
- Phenylalanine - Phe - F
- Proline - Pro - P
- Valine - Val - V

Most protein molecules have a hydrophobic core, which is not accessible to solvent and a polar surface in contact with the environment (although membrane proteins follow a different pattern). While hydrophobic amino acids build up the core of the molecule, polar and charged amino acids preferentially cover the surface and are in contact with solvent due to their ability to form hydrogen bonds. For a hydrogen bond to be formed, two electronegative atoms (for example in the case of an alpha-helix the amide N, and the carbonyl O) have to interact with the same hydrogen. The hydrogen is covalently attached to one of the atoms (called the hydrogen-bond donor), but interacts electrostatically with the other atom (the hydrogen bond acceptor, O). In proteins essentially all groups capable of forming H-bonds (both main chain and side chain, independently of whether the residues are within a secondary structure or some other type of structure) are usually H-bonded to each other or to water molecules. Due to their electronic structure, water molecules may accept 2 hydrogen bonds, and donate 2, thus being simultaneously engaged in a total of 4 hydrogen bonds. Water molecules may also be involved in the stabilization of protein structures by making hydrogen bonds with the main chain and side chain groups in proteins and even linking different protein groups to each other. In addition, water is often found to be involved in ligand binding to proteins, mediating ligand interactions with polar or charged side chain- or main chain atoms. It is useful to remember that the energy of a hydrogen bond, depending on the distance between the donor and the acceptor and the angle between them, is in the range of 2-10 kcal/mol. A detailed atlas of hydrogen bonding for all 20 amino acids in protein structures was compiled by Ian McDonald and Janet Thornton (<http://www.bmb.uga.edu/wampler/tutorial/>).

Positively and negatively charged amino acids often form so called salt bridges. These interactions may be important for the stabilization of the protein three-dimensional structure - for example proteins from thermophilic organisms (organisms that live at elevated temperatures, up to 80-90 °C, or even higher) often have an extensive network of salt bridges on their surface, which contributes to the thermostability of these proteins, preventing their denaturation at high temperatures.

The preferred location of different amino acids in protein molecules can be characterized by calculating the extent by which an amino acid is buried in the structure or exposed to solvent. Below you can see a figure showing the distribution of the different amino acids within protein molecules:

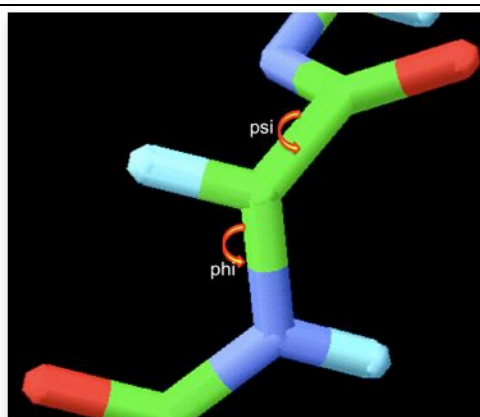


While hydrophobic amino acids are mostly buried within the core of the structure, a smaller fraction of polar groups are found to be buried. Charged residues are exposed to solvent to a much higher degree. The vertical axis shows the fraction of highly buried residues, while the horizontal axis shows the amino acid names in one-letter code. Taken from the tutorial by J.E. Wampler.

Torsion angles and the Ramachandran plot

Two **torsion angles** in the polypeptide chain, also called **Ramachandran angles** (after the Indian physicist who worked on modeling the interactions in polypeptide chains, Ramachandran, GN, et al., J Mol Biol, 7:95-99) describe the rotations of the polypeptide backbone around the bonds between N-C α (called Phi, ϕ) and C α -C (called Psi, ψ , see below for the graphics view of the angles). A special way for plotting protein torsion angles was also introduced by Ramachandran and co-authors, and was subsequently named the **Ramachandran plot**. The Ramachandran plot provides an easy way to view the distribution of torsion angles in a protein structure. It also provides an overview of excluded regions that show which rotations of the polypeptide are not allowed due to steric hindrance (collisions between atoms). The Ramachandran plot of a particular protein may also serve as an important indicator of the quality of its three-dimensional structures (*see below*).

Torsion angles are among the most important local structural parameters that control protein folding - essentially, if we would have a way to predict the Ramachandran angles for a particular protein, we would be able to predict its fold. The torsion angles provide the flexibility required for the polypeptide backbone to adopt a certain fold, since the third possible torsion angle within the protein backbone (called omega, ω) is essentially flat and fixed to 180 degrees. This is due to the partial double-bond character of the peptide bond, which restricts rotation around the C-N bond, placing two successive α -carbons and C, O, N and H between them in one plane. Thus, rotation of the protein chain can be described as rotation of the peptide bond planes relative to each other.

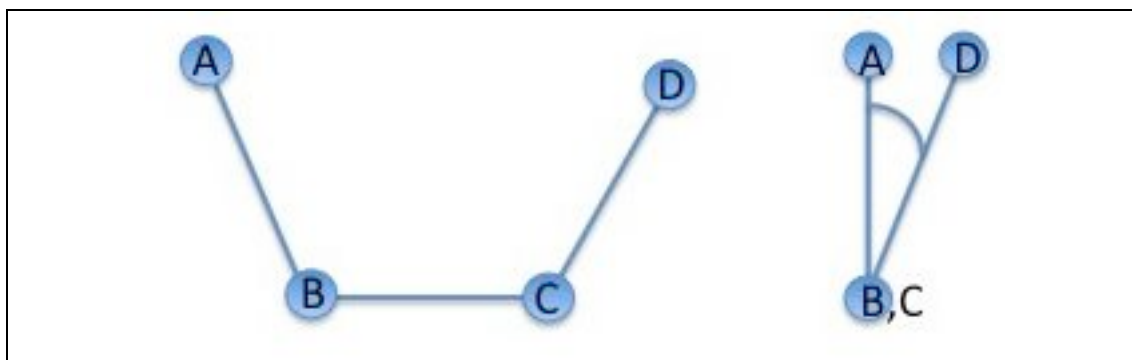


Illustration

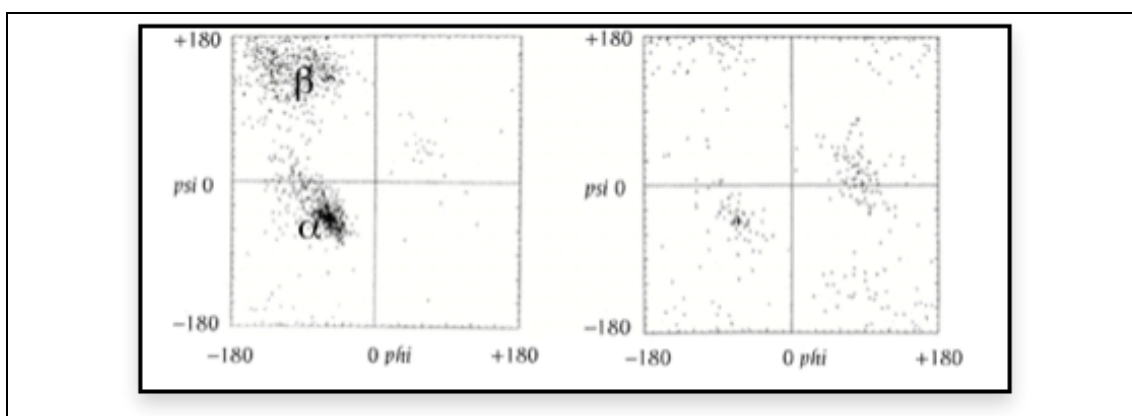
Torsion angles are dihedral angles, which are defined by 4 points in space. In proteins the two torsion angles ϕ and ψ describe the rotation of the polypeptide chain around the two bonds on both sides of the C α atom, as shown in the figure.

The standard IUPAC definition of a dihedral angle is illustrated in the figure below. A, B, C and D illustrate the position of the 4 atoms used to define the dihedral angle. The rotation takes place around the central B-C bond. The view on the right is along the B-C bond

with atom A placed at 12 o'clock. The rotation around the B-C bond is described by the A-B-D angle shown of the right figure: Positive angles correspond to clockwise rotation:



The mentioned above, restriction of the Ramachandran angles in proteins to certain values is clearly visible in the Ramachandran plot below. The plot shows that each type of secondary structure elements occupies its characteristic range of ϕ and ψ angles, marked α is for α -helices and β is for β -sheet on the left:



The horizontal axis shows ϕ values, while the vertical shows ψ values. Each dot on the plot shows the angles for an amino acid. Notice that the counting starts in the left hand corner from -180 and extend to +180 for both the vertical and horizontal axes. This is a convenient presentation and allows clear distinction of the characteristic regions of α -helices and β -sheets. The regions on the plot with the highest density of dots are the so-called “allowed” regions, also called low-energy regions. Some values of ϕ and ψ are forbidden since the involved atoms will come too close to each other, resulting in a steric clash. For a high-quality and high resolution experimental structure these regions are usually empty or almost empty - very few amino acid residues in proteins have their torsion angles within these regions. But there are sometimes exclusions from this rule - such values can be found and they most probably will result in some strain in the polypeptide chain. In such cases additional interactions will be present to stabilize such structures. They may have functional significance and may be conserved within a protein family (Pal and Chakrabarti, 2002).

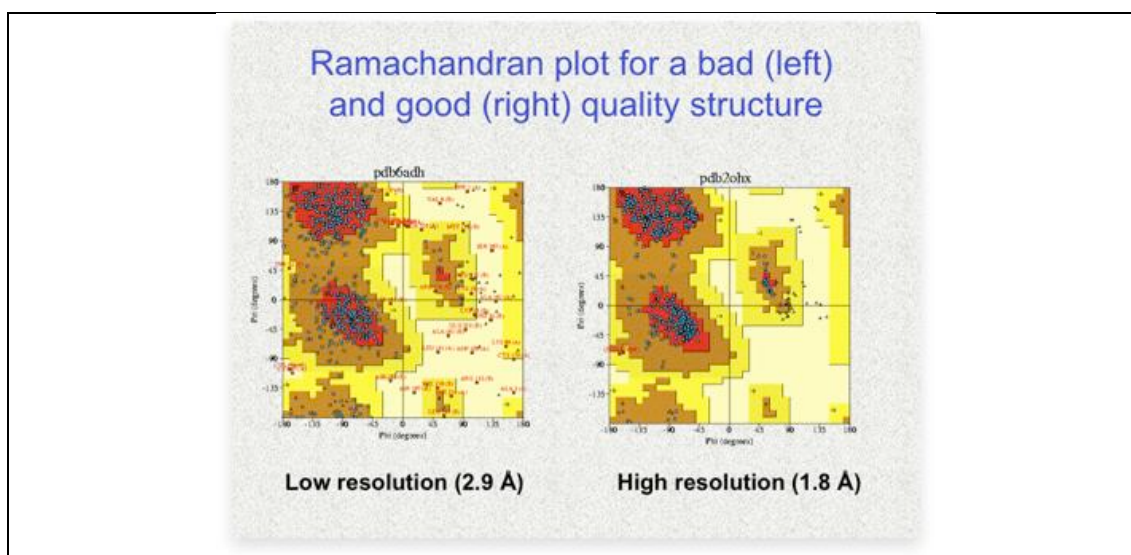
Another exception from the principle of clustering around the α - and β -regions can be seen on the right plot of the above figure. In this case the Ramachandran plot shows torsion angle distribution for one single residue, glycine. Glycine does not have a side chain, which allows high flexibility in the polypeptide chain, making otherwise forbidden rotation angles accessible. That is why glycine is often found in loop regions, where the polypeptide chain needs to make a sharp turn. This is also the reason for the high conservation of glycine residues in protein families, since the presence of turns at certain positions is a characteristic of a particular fold of a structure. Another residue with special properties is proline, which in

contrast to glycine fixes the torsion angles at a certain value, very close to that of an extended β -strand. Proline is often found at the end of helices and functions as a “helix disruptor”.

Theoretically, the average phi and psi values for α -helices and β -sheets should be clustered around -57, -47 and -80, +150, respectively. However, for real experimental structures these values were found to be different. In a paper by Hovmöller et al., 2002 in Acta Crystallographica (<http://www.fos.su.se/%7Esvenh/Conformations.pdf>), you can find detailed discussion of the fine structure of ϕ - and ψ -angle distribution in the Ramachandran plot.

The Ramachandran plot and the quality of a protein structure

In cases when the protein X-ray structure was not properly refined, and especially for bad or wrong homology models, we may find torsion angles in disallowed regions of the Ramachandran plot – this type of deviations usually indicates problems with the structure. Based on this, the Ramachandran plot is usually used in assessing the quality of experimental structures or homology models. The image below shows two Ramachandran plots for the same structure refined at different resolutions. The structure on the left was refined sometime at the early days of protein crystallography, while the one on the right was refined using more modern refinement programs. Red indicates low-energy regions, brown allowed regions, yellow the so-called generously-allowed regions and pale-yellow marks disallowed regions. On the left plot you may see many dots in the disallowed regions, but almost none on the right (the ones which are seen are for glycine residues). You may also notice that the torsion angles on the left plot lack real clustering around secondary structure regions and have a much wider distribution, compared to the plot on the right (also compare to the left plot on the figure above). Generally this is a result of bad geometry - high resolution structures generally tend to have better clustering within the allowed regions of the plot:



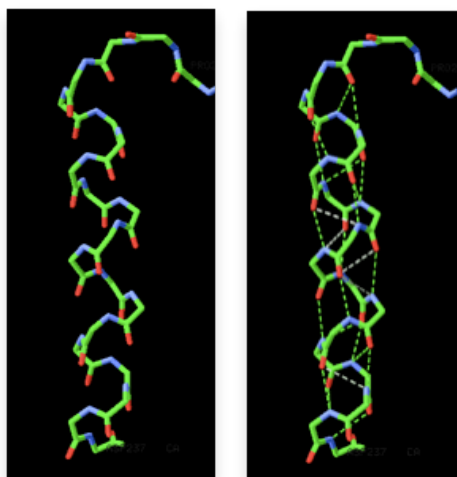
Torsion angles outside the low-energy regions, whenever observed, should be carefully examined. They may indicate problems in the structure, but they may also be true and may provide some interesting insights into the function of the protein.

Secondary structure elements

Here we will focus on the general aspects of protein secondary structure. Some features are essential in practical applications – for example in sequence alignment analysis, in

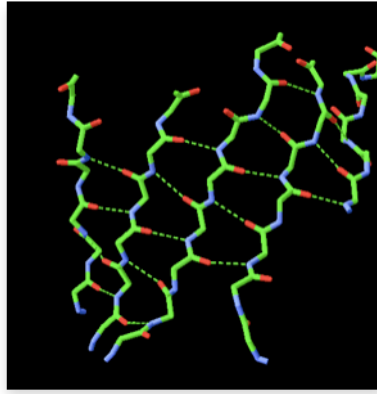
homology modeling and analysis of model quality, in planning mutations in a protein or when analyzing protein-ligand interactions.

The most common type of secondary structure in proteins is the α -helix. Linus Pauling was the first to predict the existence of α -helices. The prediction was confirmed when the first three-dimensional structure of a protein, myoglobin (by Max Perutz and John Kendrew) was determined by X-ray crystallography. An example of an α -helix is shown on the figure below. This type of representation of a protein structure is called sticks representation. To give you a better impression of how a helix looks like, only the main chain of the polypeptide is shown in the figure, no side chains. There are 3.6 residues/turn in an α -helix, which means that there is one residue every 100 degrees of rotation ($360/3.6$). Each residue is translated 1.5 Å along the helix axis, which gives a vertical distance of 5.4 Å between structurally equivalent atoms in a turn (pitch of a turn). The repeating structural pattern in helices is a result of repeating ϕ values and ψ values, observed as mentioned earlier in the text, as clustering of the corresponding torsion angles within the helical region of the Ramachandran plot. The α -helix is the major structural element in proteins. When looking at the helix in the figure below, we notice how the carbonyl oxygen atoms C=O (shown in red) point in one direction, towards the amide NH groups 4 residues away ($i, i+4$). Together these groups form a hydrogen bond, one of the main forces in the stabilization of secondary structure in proteins. The hydrogen bonds are shown on the right figure as dashed lines.



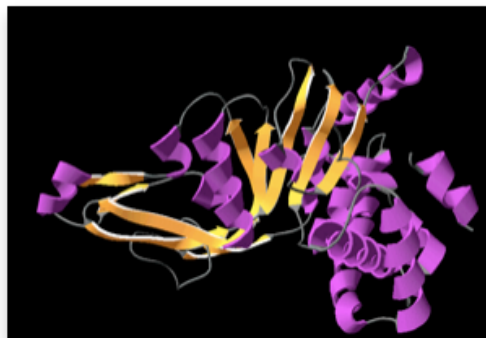
The α -helix is not the only helical structure in proteins. Other helical structures include the 3_{10} helix, which is stabilized by hydrogen bonds of the type ($i, i+3$) and the π -helix, which is stabilized by hydrogen bonds of the type ($i, i+5$). The 3_{10} helix has a smaller radius, compared to the α -helix, while the π -helix has a larger radius. A paper describing the occurrence of the π -helix in proteins, which is based on the analysis of entries in the Protein Data Bank (PDB) has been published by Fodje & Al-Karadaghi, 2002.

The second major type of secondary structure in proteins is the β -sheet. β -sheets consist of several **β -strands**, stretched segments of the polypeptide chain, kept together by a network of hydrogen bonds. An example of a β -sheet with the stabilizing hydrogen bonds shown as dashed lines is shown on the figure below:

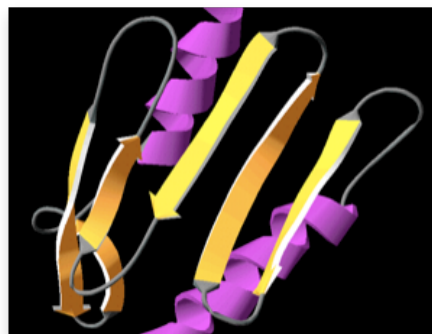


The figure shows how hydrogen bonds link different segments of the polypeptide chain. These segments do not need to follow to each other in the sequence and may be located in different regions of the polypeptide chain.

The same β -sheet is shown on the figure below, this time in the context of the 3D structure to which it belongs and in a so-called "ribbon" representation (the coloring here is according to secondary structure - β -sheets in yellow and helices in magenta). In the figure each β -strand is represented by an arrow, which defines its direction starting from the N-terminus to the C-terminus. When the strand arrows point in the same direction, we call such β -sheet **parallel**:



And when the arrows point in opposite directions, the sheet is **anti-parallel**. In the next figure you can see an example of a protein structure with an anti-parallel β -sheet:



When there are only 2 anti-parallel β -strands, like in the figure below, it is called a **β -hairpin**. The loop between the two strands is called **β -turn**, when it is short. Short turns and longer loops play an important role in protein 3D structures, connecting together strands to strands, strands to α -helices, or helices to helices. The amino acid sequences in loop regions are often highly variable within a protein family. But in some cases, when a loop has some specific function, for example interaction with another protein, the sequence may be conserved. Loop length in proteins from organisms living at elevated temperatures

(thermophilic organisms) are usually shorter than their mesophilic counterparts, presumably to give a protein additional stability at high temperatures, preventing its denaturation.



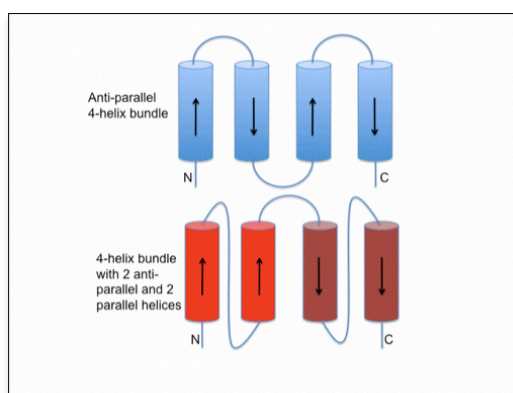
β -hairpin

You may have heard the expression "Structure is Function". This also includes various structural motifs, which are often closely linked to protein function. For this reason, when working or just viewing protein 3D structures, it is an advantage to be able to recognize the secondary structure elements and to identify structural motifs. In the next section we will look at some of the ways by which secondary structure elements may be connected to each other, forming common structural motifs. To create the observed variety of protein structures, proteins use these structural motifs as building blocks.

Structural motifs: Connectivity between secondary structure elements

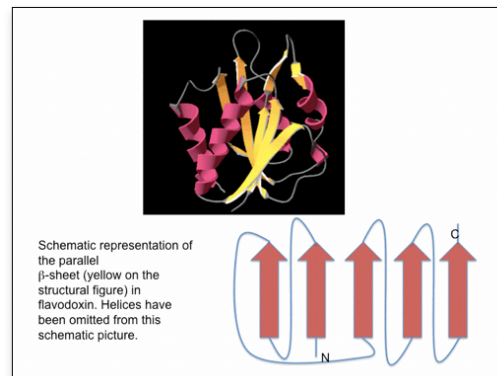
In protein structures helices and strands are connected to each other and combined in many different ways. Also, from known protein three-dimensional structures we have learned that in nature there is a limited number of ways by which secondary structure elements are combined. The connectivity between secondary structure elements and the type of secondary structure elements involved define the level of structural organization called structural motifs. Here we will look at some examples. It is possible to learn how to distinguish different structural motifs by analyzing a protein structure using graphics display software like Chimera or Pymol.

One of the simplest protein structural motifs is a helical bundle, shown on the schematic image below. Helix bundles are very common in protein structures and are very often found as separate domains within larger, multi-domain proteins.

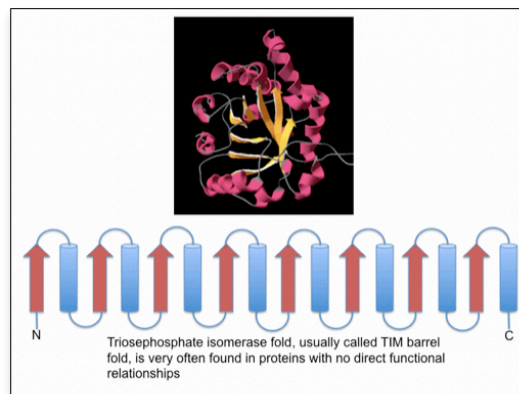


Parallel and anti-parallel β -sheets are also connected by a variety of connectivity types. The simplest and most common connectivity is made by loops, like in hairpins described earlier. If a connecting region cannot be classified as a secondary structure, and it is not a short loop, it

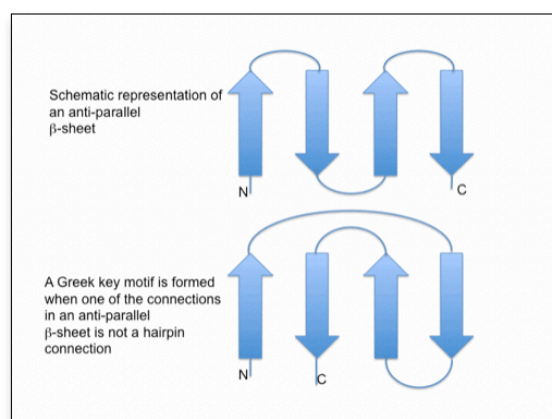
is sometimes called **coil region**. Often secondary structure elements have long coil (unstructured) regions between them. An example is shown on the figure below.



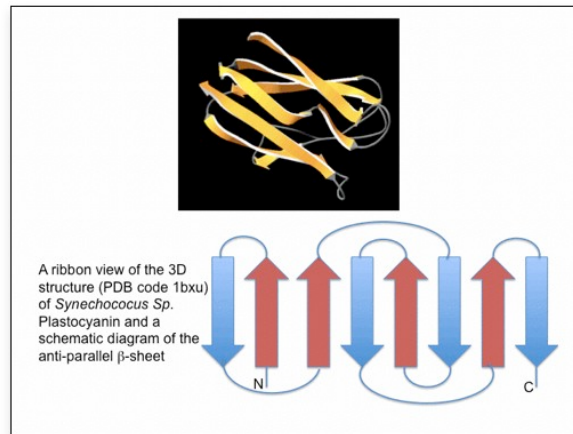
In the TIM barrel fold (the name is based on the protein where the fold was first identified, **T**riose phosphate **I**so**M**erase), the strands of the β -sheet are parallel, and the connectivity between them is made up by α -helices:



Other examples of connectivity in anti-parallel sheets are shown below. In the first two hairpins are connected to each other making up the sheet, while in the second there is the so-called Greek-key motif type of connectivity:



The figure below shows the topology of a protein plastocyanin, which only contains β -structures. Try to identify the Greek-key motif in the structure:



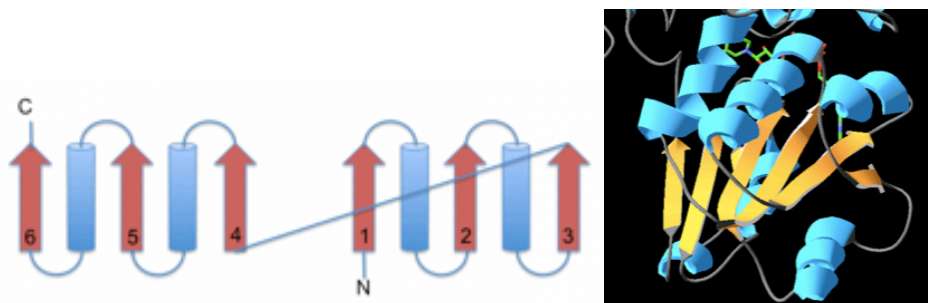
There are of course other types of connectivity between secondary structure elements. Here we just want to explain the concept by showing some example.

Folds and fold classification

Fold assignment is one of the first steps in the analysis of protein structure. Fold analysis may reveal evolutionary relationships, which sometimes are difficult to detect at the sequence level, it may also help a better understanding of the mechanism of function of a protein, its activity and biological role. Study of the relationships between the amino acid sequence and the fold may also reveal deeper insights into the fundamental principles of **protein structure**, and may aid, e.g. in the design of new proteins with pre-defined structure and activity.

The relationship between the amino acid sequence and the three-dimensional structure of a protein is not unique – a large number of modifications in the sequence within a protein family can be tolerated and will result in a similar 3D structure. The higher degree of conservation of the three-dimensional structure, compared to sequence conservation, is a prerequisite for the function of a protein (structure is function!). By other words, the constraints put during evolution by Nature on the three-dimensional structure are much tighter than those put on the amino acid sequence. There are special techniques used to compare 3D structures and to judge the degree of similarity between them. Some discussion on this subject may be found in the homology modeling chapter.

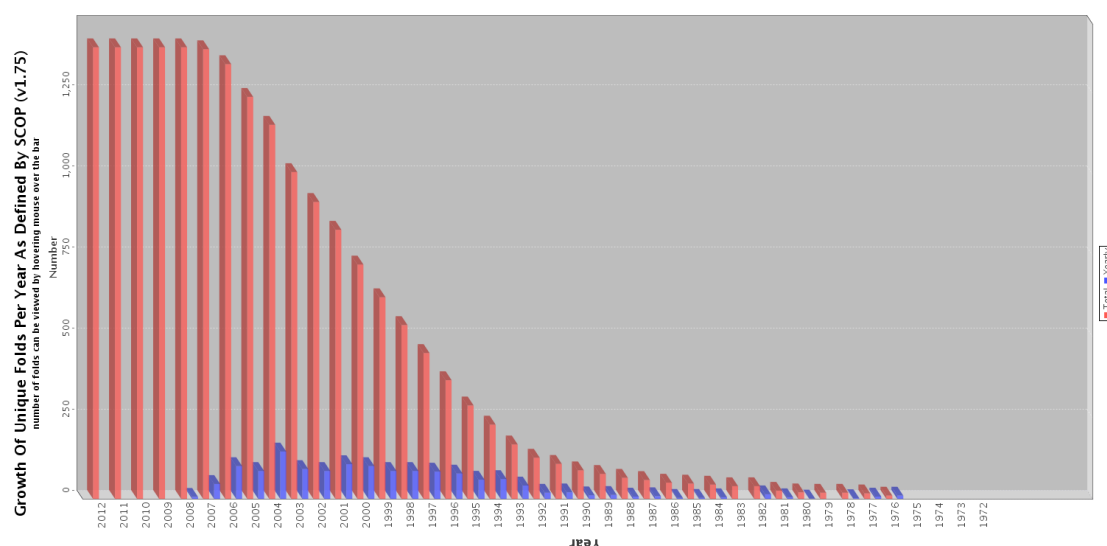
A **protein fold** is defined by the arrangement of the secondary structure elements of the structure relative to each other in space. Some folds have already been mentioned in the previous section on protein motifs. The 4-helix bundle and the TIM barrel, for example, are two types of very common protein folds. The amino acid sequences of proteins forming these two folds may lack any evolutionary relationships, still producing similar 3D structures. An additional example is shown below. It is the coenzyme-binding domain of some dehydrogenases, which adopts the so called Rossmann fold, named after Michael G. Rossmann, a protein crystallographer who solved one of the very first structures with this type of fold. It is also the only protein fold named after the person who was first to discover it:



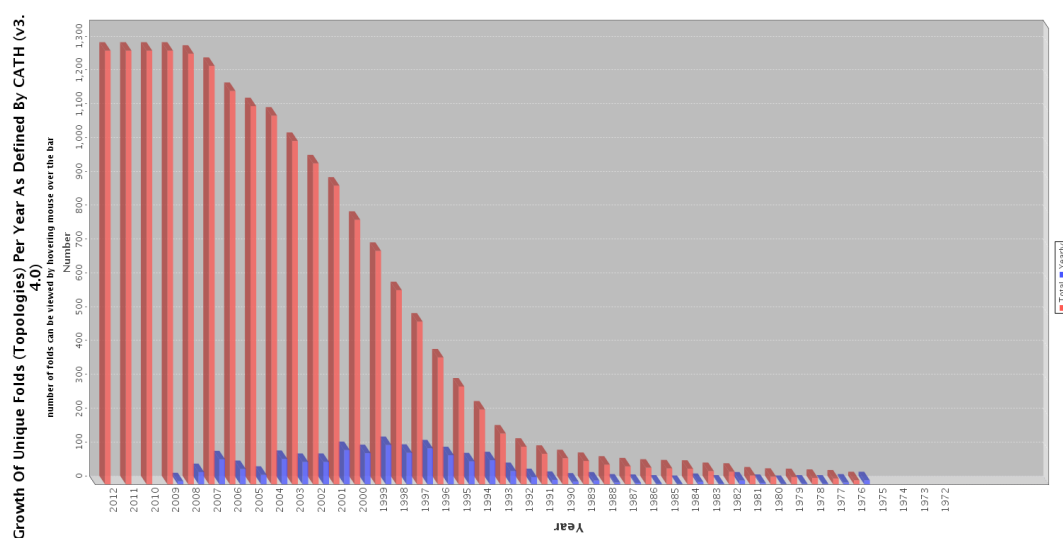
In this figure on the left a schematic presentation of the Rossmann fold. On the right the nucleotide binding domain of liver alcohol dehydrogenase is shown. Notice the central parallel β -sheet (shown in yellow) flanked by α -helices on both sides of its plane. There are of course many more types of protein folds, but how many in total? Taking into account the huge number of amino acid sequences, one would expect a high number of different folds. But in reality the number of folds is limited. Nature has re-used the same fold again and again for performing totally new functions. To check statistics on protein folds we can simply go to the Protein Databank (PDB) and click the PDB Statistics link on the right upper corner. This will bring us to a page where among other stuff the following two options are shown:

- **Folds As Defined By SCOP**
- **Topologies As Defined By CATH**

SCOP and CATH are the two databases generally accepted as the two main authorities in the world of fold classification. According to SCOP there are **1393** different folds. Also notice the graph, the last time a new fold was identified was 2008:



The next graph shows the folds identified by CATH database, a total of **1282** folds:



Apparently the two databases use slightly different ways for fold definitions and classification, which results in different total numbers of folds. It is also interesting to note that during the recent years essentially no new folds have been discovered. Have we reached

the limit? There is probably still a chance that some new folds will be discovered.

Since many proteins contain several domains with different folds, one could ask: What is actually being classified by these databases? The answer is the "simplest", or sometimes also called the "independent" folding unit of a protein – a domain. Knowing the fold of the different domains in a protein molecule is important in many cases. For example, in homology modeling we need to have a clear idea about the number of domains in a protein and the type of folds they have.

Domains and domain classification

Many proteins only contain a single domain, while others may have several domains. Some domains have some clearly defined function associated with them, like the Rossmann-fold domain, also called coenzyme-binding domain, discussed earlier. Such domains often “carry” their function with them when they get inserted into different proteins during evolution. A domain may be characterized by the following:

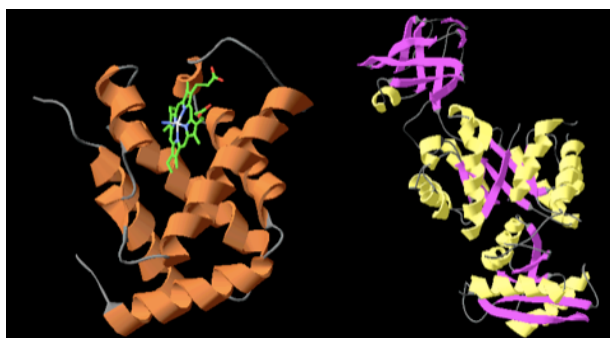
- 1- A spatially separated unit of the protein structure
- 2- Often have sequence and/or structural resemblance to some protein structure or domain.
- 3- Often have a specific function associated with it.

The easiest way to follow for the characterization of the fold of a protein domain would be to search in the respective databases. The procedure followed by databases, for example CATH or SCOP, includes:

- 1- Assignment of secondary structure
- 2- Assignment of domains
- 3- Assignment of a structural class to each domain (3 possible structural classes, alpha, beta and alpha/beta)
- 4- Assignment of fold (called Architecture in the CATH database)
- 5- Assignment of topology (homologues superfamily)

Secondary structure is usually assigned automatically, using computer software. All protein structure visualization programs like Chimera and Pymol include this function, and all PDB files contain definition of secondary structure in a protein (shown in beginning of the file).





One needs to be aware that CATH and SCOP use slightly different terminology in fold assignment and have a different way of describing the entries. CATH follows the Class-Architecture-Topology-Homologous superfamily classification scheme. There are currently 53 million protein domains classified into 2,737 superfamilies in the CATH database. As an example, the figure below shows two proteins, one contains one domain (hemoglobin), while the second has 3 domains (pyruvate kinase). A subunit of hemoglobin consists of a single α -helical domain. You may also see the heme molecule (in sticks representation) bound within a pocket created by the α -helices:



[illegible]

Kinase. This information is highly valuable in homology modeling, especially in cases when we need to model different domains using different modeling templates, the so called multi-template homology modeling.

CATH Classification

Level	CATH Code	Description
	3	Alpha Beta
	3.40	3-Layer(aba) Sandwich
	3.40.1380	Pyruvate Kinase; Chain: A, domain 1
	3.40.1380.20	

CATH Clusters

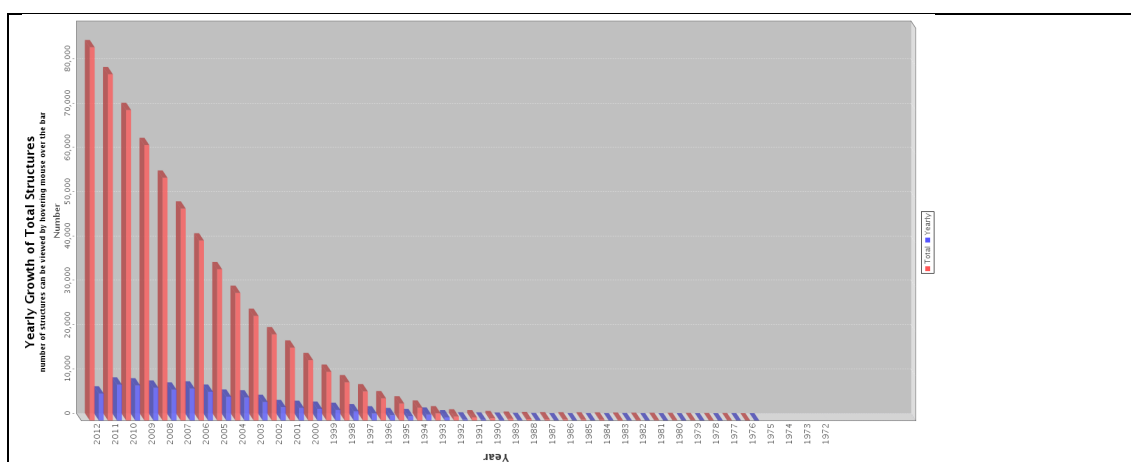
Superfamily	3.40.1380.20
Functional Family	Pyruvate kinase -like domain
Structural Cluster	SSG

In the next section we will look at the PDB and PDBsum protein databases, which are going to be used later in the homology modeling project.

Protein databases: PDB & PDBsum

There are many protein and structural bioinformatics-related resources on the Internet. Some of them are of general character, some are dedicated to specific aspects of protein families, specific metabolic pathways, etc. Here we will discuss just few general-character databases.

The first question, when working with a protein, would be where to find its structure. It is also interesting to know what is actually inside a structural file, what type of information is kept there and how structural information is presented in the file. The primary database for protein structure information is the **Protein Data Bank (PDB)**, created sometime in the beginning of the 1970ties. Only few structures existed at that time, and the only experimental method for protein structure determination available was protein X-ray crystallography. The real structural revolution, started in the 1990ties:



One of the reasons for this structural revolution was that cloning techniques started to enter

the lab and both the number and amount of proteins available for crystallization increased drastically. Before the cloning era people had to purify proteins from cells, substantially limiting availability – to obtain a few milligrams of a protein for crystallization one would need a lot of cells. Cloning solved the problem, proteins could be expressed in large quantities and purified for crystallization. Another important factor was the introduction of synchrotron radiation. Synchrotrons, like MAX IV in Lund, Sweden, ESRF in Grenoble, France, or DESY in Hamburg, Germany, and many others around the world provide very high intensity X-rays, which may be used for collecting high quality X-ray diffraction data even from small crystals. This eliminated the time-consuming stage of optimization of the crystallization conditions, which was required for obtaining crystals large enough for the relatively low X-ray intensity of home sources. The third factor was probably the introduction of personal computers, relatively cheap and with ever increasing power. Cheaper computers also meant new software, which also started to become user friendly, and in addition new graphics capabilities of monitors became available. A proper graphics monitor with a computer, which was used for model building in the early days of crystallography would cost around 50-60 thousands dollar! Now a better PC or a Mac is all we need. That was when the number of protein structures started to increase dramatically. Then came the era of **structural genomics** – large consortia were formed with the aim to develop new technology for solving large amounts of protein structures. One such consortium is, for example, the Structural Genomics Consortium (SGC). With the increasing number of structures the number of protein databases started to increase and new tools for the analysis of protein sequence and structure were rapidly developed.

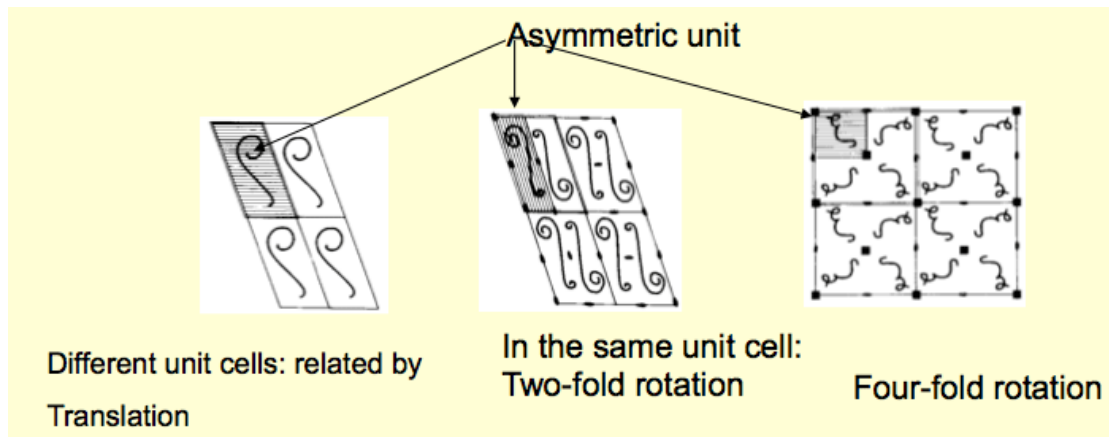
Currently every newly determined protein structure has to be deposited with the Protein Data Bank before the scientific paper describing the structure can be published. Currently the number of structures in the PDB has exceeded 100 000. However, one should remember that not all structures in the PDB are unique. In many cases there are many entries of the same protein in the database – some are mutant variants, others may be complexes with ligands (substrate analogues, inhibitors, co-factors), complexes with other proteins, etc. This may be a source of confusion if one would try to fetch a structure from PDB – which one to choose if there are many entries of the same protein? This will be discussed later in the chapter on homology modeling. For modeling it is important to choose the right structure with the best available quality.

Coming back to our initial questions, how to download a structure and what is inside the PDB file? First we need to check if there is a structure for the protein we are interested in. This part is easily done, all you need to do is to go to the PDB and type the name of the protein you are looking for into the search window. For example, enter the name of a protein called magnesium chelatase. Generally one would get several hits, however, in the case of magnesium chelatase there is only one X-ray structure for one of the submits of the enzyme. Some other proteins may be listed in the output, some of them come from electron microscopy modeling, others may be totally unrelated. PDBsum gives more clear results – entering the name of the same protein we would get a single hit (PDB ID 1g8p). Of course you may refine your search using the options provided on the PDB page that show up when you enter the name of the protein. Among the options to refine our search we can choose the organism from which the protein originates, chose a particular subunit, the experimental method, etc.

Both PDB and PDBsum provide additional data on the entry, including links to other databases, where more information can be found. Here is an example from PDBsum link page:

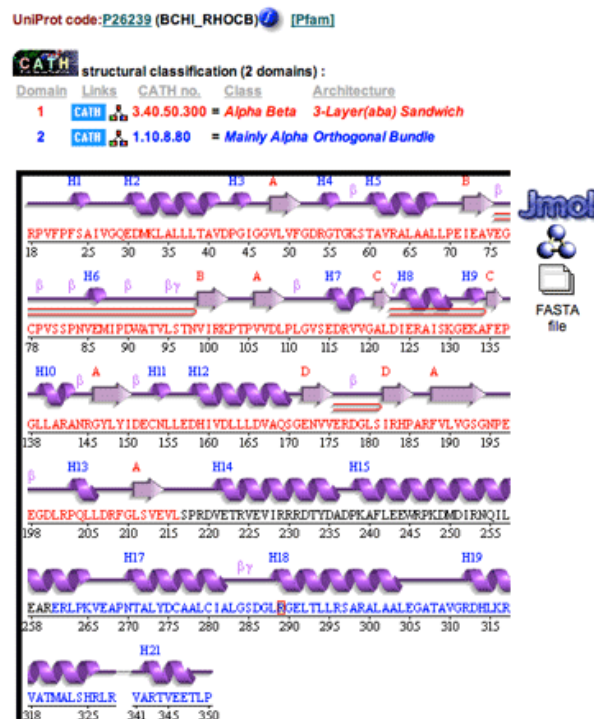
Links to other databases	
Structure databases	
PDBe	Protein Data Bank, Europe at the EBI
RCSB	Protein Data Bank at the RCSB
SRS	SRS at the EBI
MMDB	MMDB entry at the NCBI
JenaLib	Jena Library of Biological Macromolecules at the Fritz-Lipman Institute
OCA	OCA at the Weizmann Institute
Wikis	
PDBWiki	PDBWiki - a community annotated knowledge base of biological molecular structures
PROTEOPEDIA	Proteopedia - collaborative 3D encyclopedia of proteins and other molecules
Fold databases	
CATH	CATH structural classification
SCOP	SCOP structural classification
FSSP	FSSP structural alignments
Protein sequence	
PDBSWS	PDBSWS: mapping of PDB protein chains to SwissProt entries
Secondary structure	
HSSP	HSSP: Homology derived Secondary Structure of Proteins
Quaternary structure	
PQS	Protein Quaternary Structure server at the EBI
Experimental data	
EDS	EDS: Uppsala Electron Density Server
Functional annotation	
CSA	CSA: Catalytic Site Atlas
ProSAT	ProSAT: Protein Structure Annotation Tool
Quality assessment	
PROCHECK	PROCHECK summary of protein structural quality
WHATCHECK	WHATIF report on protein structural quality

For our purposes we may be interested in the links to CATH and SCOP (structural classification). The PQS database is also of interest, it is the Protein Quaternary Structure database. However, when you click on this link the database will inform you that from 2009 it is not updated anymore. The reason is that the information, which can be found in PQS is currently generated by the PISA sever, **P**rotein **I**nterfaces, **S**urfaces and **A**ssemblies. The reason is that PDB files usually contain the crystallographic unit, the so-called asymmetric unit. The biological unit in solution may contain several subunits of the same protein, arranged as dimers, trimers or higher order oligomers. In these oligomers the subunits are usually related by some kind of symmetry - two-fold rotation for dimers, three-fold rotation for trimers, four-fold rotation for tetramers, etc. When the molecules are crystallized, they get arranged in certain types of space lattices, within which all molecules are ordered and related to each other by symmetry operations of the particular symmetry group of the crystal (possible symmetry groups are listed in the International Tables for Crystallography). The symmetry axes present in the molecule in solution, which could be 2-, 3-, or 4-fold, may become part of the crystallographic symmetry. In such cases, one unit within, for example a trimer, becomes the asymmetric unit of the crystal. Crystallography operates with asymmetric units since the other units will be exactly the same and related by the symmetry operation of the crystal. This is reflected in the content of the files in the PDB, they contain coordinates for the atoms of one subunit, the asymmetric unit. The PISA server reconstructs the biological unit in cases when it is known to be different from the asymmetric unit or when there are some other indications that need to be taken into account. The file generated by the PISA server may also be downloaded from the PDB. The concept of the asymmetric unit is illustrated in the figure below:



In the left figure the asymmetric unit of the crystal is just one subunit and all molecules in the lattice are related to each other by simple translation. In the middle figure there are two subunits in the unit cell related to each other by a two-fold rotation symmetry axis. In this case there is a big chance that the biological unit of the protein in solution is a dimer. In the last figure on the right the molecules in the unit cell are related by a 4-fold crystallographic symmetry axis. Again, it cannot be excluded that the biological unit is going to be a tetramer.

One essential feature is a description of the amino acid sequence in relation to the secondary structure of the protein. This is provided both by PDB and PDBsum. The image below shows the page from PDBsum:



The information in this page is useful for quick identification of the position of amino acids within the structure, for getting an idea on the type of the protein (all α , α/β), the location of active site residues, etc. There is also a reference to the publication that describes the structure (rather detailed in PDBsum, even with links to citing papers).

The Protein Databank (PDB): File Format and Content

Here we will focus on the PDB. We could start with downloading the coordinate file, opening it in some text editor to look into its content. PDB files are simple text files and can be open by any text editor (in contrast, for example to MS Word files, which cannot be opened by all text editors). The file is called a "coordinate file" simply because it contains a list of the coordinates of all the atoms of the protein structure in some conventional orthogonal coordinate system (at least the atoms visible in the electron density map calculated on the basis of the X-ray data). Each atom position is defined by its x,y,z coordinates. To download the file, we simply search first by typing the name of the protein, and when we find the entry we are interested in, like the magnesium chelatase protein we discussed briefly earlier, we can open the drop-down menu in the right corner, as shown on the image below, choose and save the file on the hard drive (easiest to choose PDB file (Text)):

The screenshot shows the PDB entry page for 1G8P. The title is "CRYSTAL STRUCTURE OF BCHI SUBUNIT OF MAGNESIUM CHELATASE". The DOI is 10.2210/pdb1g8p/pdb. The primary citation is from Fodje, M.N., Hansson, A., Hansson, M., Olsen, J.G., Gough, S., Willows, R.D., Al-Karadaghi, S., published in J. Mol. Biol. 311: 111-122 (2001). The PubMed ID is 11469861. The abstract describes the structure of the BchI subunit of magnesium chelatase. On the right, a dropdown menu is open, showing options to download the file in various formats: FASTA Sequence, PDB File (Text), PDB File (gz), mmCIF File, mmCIF File (gz), PDBML/XML File, PDBML/XML File (gz), Structure Factor (Text), Structure Factor (gz), and Biological Assembly (gz) (A).

There is a plenty of important information on the structure in the file, like the method used to solve the structure and various parameters related to the quality of the X-ray data (like resolution, R-factor etc.) and the structure as such, like geometry, secondary structure content, regions missing in the structure, etc). The R-factor is an essential parameter for the assessment of how well the structure fits the X-ray data. The lower the value of the R-factor, the better the fit. Well-refined protein structures have R-factor values below 20%:

```
REMARK 1
REMARK 2
REMARK 2 RESOLUTION.      2.10  ANGSTROMS.
REMARK 3
REMARK 3 REFINEMENT.
REMARK 3   PROGRAM          : CNS 1.0
REMARK 3   AUTHORS         : BRUNGER,ADAMS,CLORE,DELANO,GROS,GROSSE-
REMARK 3                   : KUNSTLEVE,JIANG,KUSZEWSKI,NILGES, PANNU,
REMARK 3                   : READ,RICE,SIMONSON,WARREN
REMARK 3
REMARK 3 REFINEMENT TARGET : ENGH & HUBER
REMARK 3
REMARK 3 DATA USED IN REFINEMENT.
REMARK 3 RESOLUTION RANGE HIGH (ANGSTROMS) : 2.10
REMARK 3 RESOLUTION RANGE LOW  (ANGSTROMS) : 29.55
REMARK 3 DATA CUTOFF              (SIGMA(F)) : 0.000
REMARK 3 DATA CUTOFF HIGH        (ABS(F))   : 312841.620
REMARK 3 DATA CUTOFF LOW         (ABS(F))   : 0.0000
REMARK 3 COMPLETENESS (WORKING+TEST) (%)    : 97.9
REMARK 3 NUMBER OF REFLECTIONS              : 22179
REMARK 3
REMARK 3 FIT TO DATA USED IN REFINEMENT.
REMARK 3 CROSS-VALIDATION METHOD              : THROUGHOUT
REMARK 3 FREE R VALUE TEST SET SELECTION      : RANDOM
REMARK 3 R VALUE                             (WORKING SET) : 0.214
REMARK 3 FREE R VALUE                       : 0.247
REMARK 3 FREE R VALUE TEST SET SIZE (%)     : 10.000
REMARK 3 FREE R VALUE TEST SET COUNT        : 2207
REMARK 3 ESTIMATED ERROR OF FREE R VALUE    : 0.005
REMARK 3
```


Further down there is a list of the secondary structure elements within the structure, also showing the first and last residue in each element:

```

HELIX  1  1 PRO A  22  ILE A  26  5  5
HELIX  2  2 GLN A  29  ASP A  42  1  14
HELIX  3  3 PRO A  43  GLY A  46  5  4
HELIX  4  4 ASP A  53  GLY A  57  5  5
HELIX  5  5 SER A  59  LEU A  69  1  11
HELIX  6  6 ASN A  84  ILE A  88  5  5
HELIX  7  7 SER A  114 GLY A  120 1  7
HELIX  8  8 ASP A  123 GLY A  131 1  9
HELIX  9  9 GLY A  138 ASN A  144 1  7
HELIX 10 10 GLU A  152 LEU A  156 5  5
HELIX 11 11 GLU A  157 GLY A  171 1  15
HELIX 12 12 ARG A  202 ASP A  207 1  6
HELIX 13 13 ASP A  220 ASP A  237 1  18
HELIX 14 14 ASP A  237 LEU A  263 1  27
HELIX 15 15 PRO A  264 VAL A  266 5  3
HELIX 16 16 PRO A  269 LEU A  283 1  15
HELIX 17 17 GLY A  287 GLU A  305 1  19
HELIX 18 18 GLY A  311 SER A  324 1  14
HELIX 19 19 HIS A  325 LEU A  327 5  3
HELIX 20 20 VAL A  341 LEU A  349 1  9
SHEET  1  A 5 VAL A 106 LEU A 109 0
SHEET  2  A 5 GLY A 146 ILE A 150 1 O TYR A 147 N VAL A 107
SHEET  3  A 5 PHE A 188 GLY A 194 1 O VAL A 189 N LEU A 148
SHEET  4  A 5 VAL A  48 PHE A  51 1 N VAL A  48 O LEU A 190
SHEET  5  A 5 LEU A 211 GLU A 214 1 O LEU A 211 N LEU A  49
SHEET  1  B 2 ILE A  72 VAL A  75 0
SHEET  2  B 2 VAL A  99 LYS A 102 -1 N ILE A 100 O ALA A  74
SHEET  1  C 2 ALA A 121 LEU A 122 0
SHEET  2  C 2 PHE A 135 GLU A 136 -1 N GLU A 136 O ALA A 121
SHEET  1  D 2 GLU A 172 VAL A 175 0
SHEET  2  D 2 ILE A 182 PRO A 185 -1 O ILE A 182 N VAL A 175
CRYST1 90.259 90.259 83.716 90.00 90.00 120.00 P 65 6

```

After the general informational part, the x,y,z coordinates of the atoms are listed:

```

ATOM  1  N  ARG A 18  14.699 61.369 62.050 1.00 39.19  N
ATOM  2  CA ARG A 18  14.500 62.241 60.856 1.00 38.35  C
ATOM  3  C  ARG A 18  13.762 61.516 59.729 1.00 36.05  C
ATOM  4  O  ARG A 18  14.354 60.740 58.982 1.00 34.91  O
ATOM  5  CB ARG A 18  15.850 62.753 60.334 1.00 42.36  C
ATOM  6  CG ARG A 18  16.537 63.770 61.247 1.00 46.92  C
ATOM  7  CD ARG A 18  17.825 64.314 60.629 1.00 51.24  C
ATOM  8  NE ARG A 18  18.442 65.347 61.462 1.00 54.15  N

```

First of all, notice that this structure starts from amino acid Arg 18! No amino acids from 1 to 17. The reason is that there was no electron density for these residues (see for example the discussion on structure quality in homology modeling). This is normally a result of a high flexibility of that particular region of the structure. It is essentially impossible to find the correct positions for amino acids without the guiding electron density. We need to be aware that many structures in the PDB have missing parts, sometimes in loop regions, sometimes just a side chain, and in the worse cases a whole domain may be missing.

The numbers after the first record in the file, ATOM, are just sequential numbers of the atoms in the structure. This is followed by the atom type - for example, CA means C- α , the carbon atom to which the side chain of the amino acid is attached. The next carbon atom is C- β , and following atoms are named after the Greek alphabet, gamma, delta, etc. Except C- α , main chain atoms do not have any Greek letters attached to them. They are just C, O and N. After the atom type, you will see the name of the amino acid, followed in this file by a letter A. This is the so-called chain identifier. In cases when the structure consists of several polypeptide chains (a **multi-subunit protein**), each chain will get its own identifier, like A, B, C, etc (as in the case of Pyruvate kinase discussed earlier). Without chain identifiers graphics programs will get confused having the same amino acids names and numbers for different chains (in cases of homo-multimeric proteins). The 3 numbers which follow (e.g., 14.699, 61.369, 62.050 for the very first atom) are the x,y,z coordinates of the atom. They describe the position of each atom in an orthogonal coordinate system. If we can describe the position of each atom in the protein, we will obviously be able to draw the whole tertiary structure. Graphics programs, when they read the coordinates from the protein databank file,

simply connect the atoms to each other according to some distance cut-offs, thus creating the graphics view we are accustomed to. For example, we know that C-C distance is 1.54 Å and this can be used to connect two carbon atoms when they are found to be at this distance from each other.

The x,y,z coordinates are followed by a number, which is one in most cases. This is called atom occupancy. Sometimes the side chain of a particular amino acid, but even main chain atoms, may have two or more different conformations due to local flexibility. These conformations can be distinguished in the electron density map of the structure. In this case the crystallographer will build both conformations into the electron density and refine a parameter called occupancy, for each conformation. In protein databank files these conformations are called "alternative conformations" and often marked with "ALT". The occupancy numbers for each alternative conformation will be less than 1 (1 corresponds to 100% occupancy), for example it may be 0.5/0.5 (50/50), when both conformations are equally occupied, or 40/60, or some other numbers. Also ligands and metal atoms bound to proteins may often have partial occupancy, for example if the concentration of the ligand or metal, which was co-crystallized with the protein or soaked into the protein crystal, was too low.

The numbers in the last column in the file are called the temperature factors, or B-factor, for each atom in the structure. The B-factor describes the displacement of the atomic positions from an average (mean) value. For example, the more flexible an atom is the larger the displacement from the mean position will be (mean-squares displacement). In graphics programs we can usually color a protein according to B-factor value. Areas with high B-factors are often colored red (hot), while low B-factors are colored blue (cold). An inspection of a protein databank structure with such coloring scheme will immediately reveal regions with high flexibility in the tertiary structure of the protein. The values of the B-factors are normally between 15 to 30 (sq. Angstroms), but often much higher than 30 for flexible regions.

Sequence Alignment and Analysis

Overview

Amino acid sequence alignment and analysis is central to most biochemical and molecular biology applications. Although it should be possible to retrieve all the information we need about a protein directly from its sequence, looking at a sequence without prior knowledge and experience is like reading a text in a foreign language: we may recognize the letters, but we do not understand the meaning and are unable to extract the information. Still, when proteins are concerned, we have learned to extract a substantial part of the information from detailed sequence analysis, using for example **multiple sequence alignment**. In a multiple sequence alignment a given sequence is compared to a group of other sequences from related organisms. When we say "related" we mean "evolutionary related" and that they belong to the same family, the members of which usually perform a similar function in different organisms. We know that when proteins are evolutionary related the main characteristic features of the sequence and the tertiary structure are conserved. Since conservation of function normally assumes that a certain number of **amino acid residues** within a protein family are conserved, we need to have some tools to be able to assess the degree of conservation of each member of the protein family. For this, alignment techniques and **scoring schemes for sequence alignment** have been developed. Here we will discuss the basic concepts behind these techniques and will provide some examples to guide you in making **sequence alignment** using Internet resources. Since we focus on structural bioinformatics, we will also need to learn how to interpret sequence alignments in terms of the three-dimensional structure of the protein, and to relate sequence and structural information. We may even use available structural data to make better sequence alignment!

In a sequence alignment we try to align identical amino acids in the sequences against each other. However, since normally there are also many amino acid substitutions, we need to know how to handle substitutions of one amino acid by another in the sequences being aligned (**amino acid substitutions are caused by mutations in the gene coding for the protein in question**). Some substitutions are conservative, i.e., they will not cause any substantial disturbances in the protein structure, which would affect the protein function, but other substitutions, if they would occur, may have a dramatic effect on protein structure and function. To handle amino acid substitutions in sequence alignment, specially designed substitution matrices are used, which are part of the alignment **scoring scheme** and help in calculating the score of the alignment to distinguish between several possible alignments. Even structural information may be used in making a correct alignment, for example in correctly placing insertion and deletion regions in the alignment. Insertions and deletions are very common in sequences belonging to the same family and often occur in loop regions. By other words, insertions and deletions may indicate that a certain region of the sequence may have a loop structure.

Sequence alignment basics

Since evolutionary relationships assume that a certain number of the **amino acid residues** in a protein sequence are conserved, the simplest way to assess the relationships between two sequences would be to count the numbers of identical and similar amino acids. This is done by sequence alignment. The number of identical and similar amino acid residues may then be compared to the total number of amino acids in the protein. This gives the percentage of identical and similar residues – percentage of sequence identity and sequence similarity. Similar residues are those that have similar chemical characteristics, like positively charged Lys and Arg, or hydrophobic Leu and Val, etc. Substitution of amino acids by chemically equivalent ones often does not have a dramatic effect on the structure or function of the protein. For example, Leu and Val will be equally tolerated within a hydrophobic core, assuming that there is place for the slightly larger side chain of leucine. The same applies to

Lys and Arg, which are usually located on the surface and primarily interact with solvent or with the acidic side chains of Glu or Asp. On the other hand, a substitution of Val by Arg may have a dramatic effect and may destabilize and even denature a protein.

To count the number of identities and similarities in sequence alignment, we need to establish some rules describing how alignment can be performed. Apparently we want to align as many identical or similar amino acid residues against each other as possible. Nevertheless, one should be aware that **an alignment generated by a computer program represents only one of many possibilities**. One of the reasons is that while identical amino acids are easy to recognize and align, alignment of similar amino acids is not that straight forward. For example, how to score and prioritize the following substitutions - Val-Leu, Leu-Ile, Ser-Thr or Lys-Arg? Apparently, the score we give to each of these substitutions, or call it a weight, may affect the entire alignment.

Additional factors to take into account when analyzing sequences are insertions and deletions – it is quite common that within a protein family some of the sequences have extra inserted (insertions), or missing residues (deletions). This can often be seen, for example, when a group of bacterial sequences is compared to a group of eukaryotic sequences. Sometimes even larger segments or a whole domain may be inserted into or deleted from a protein. Depending on how we handle these insertions and deletions, different sequence alignments may be generated. To illustrate the concept, an example of a simple alignment of a short stretch of two sequences is shown below. This was extracted from a ClustalW generated sequence alignment using the EBI server (European Bioinformatics Institute):

```
1: GCPVS-SPNVEM
2: GCPYGCDFEADA
   GCPxx-xPxxxx
```

The amino acids that are identical (conserved) in the two sequences are marked in the third row by their names (GCP and P), while those which are different are marked by x. You may also see that one of the cysteine in the second sequence does not seem to have a corresponding mate in the first. This position is marked by a dash. The percentage of identity for this sequence alignment is simply 4/12, or 30%. Then, the score of the alignment can be assessed, for example, by a simple expression:

$$(\text{Score}) S = \text{number of matches} - \text{number of mismatches} = 4 - 12 = -8$$

Everything looks fine, except that to maximize the number of matches, we introduced a **gap** (marked by a dash in the first sequence). A gap in one of the sequences simply means that one or more amino acids have been deleted from the sequence, or we could also say that there is an insertion in the second sequence. When introducing a gap several questions may arise: How many gaps can we introduce? How to decide where to place them? How long they can be? By introducing a large number of gaps here and there, we could continue maximizing the percentage identity, but would that be biologically relevant? Intuitively one would think that this couldn't be correct simply because behind each structure there is a three-dimensional structure and a structure can be easily disrupted by a large number of insertions and deletions. For example in homology modeling an incorrectly placed may result in a totally meaningless model. Normally, when we run sequence alignment software, we will notice that the number of gaps is limited - the alignment program must have some instructions on how to limit the number of gaps and where to place them. These instructions are **gap penalties**. Each time the program introduces a gap it triggers a penalty score, which reduces the total score of the alignment. However, this would make the whole thing meaningless, unless gap introduction would rise the total score by a value that is higher than negative effect of the penalty. By this simple way we can limit the number of gaps and increase their significance. The value of gap

penalties is a parameter that can be changed during each time an alignment is run. This will affect the number of gaps, their length and position in the sequence alignment.



Amino acid substitutions and amino acid replacement matrices (PAM, PET91, BLOSUM)

Imagine a deletion of a couple of residues, for example, in an α -helix. What will happen to that helix? There is a good chance that it will change its shape or even collapse, since a deletion or insertion will introduce a distortion in the hydrogen bonding network, in the packing of side chains, in the mutual adjustment of the torsion angles along the helix, etc. This in turn, may modify the overall 3D structure of the protein, affecting its function, or probably resulting in denaturation and total loss of function, loss of the protein's ability to interact with its partners, etc. For this reason, insertions and deletions are usually found in regions between secondary structure elements – loop regions, where they can be accommodated easier without major distortions in the overall fold of the protein. The core of proteins, on the other hand, normally has a higher degree of sequence conservation and a smaller number of insertions and deletions, which is reflected by a smaller number of gaps in such regions of the sequence alignment.

Generally, we cannot score the alignment only according to the number of aligned identical and similar residues, we also need to take into account the number of gaps in the alignment, their length and position in the sequence. To do this, various types of gap penalties are introduced – gap opening penalty, gap extension penalty, etc. Then, we need to allow gaps only at positions where they would increase the total score of the alignment even after taking into account the imposed penalties:

$$S = \Sigma \text{ of costs (identities, replacements)} - \Sigma \text{ of penalties (number of gaps} \times \text{gap creation penalties)}$$

The numbers for identities and replacements used for calculating the overall alignment score in the expression above are usually presented in the form of a 20 x 20 matrix (20 is the number of the most common amino acids). In total there are 210 possible replacement pairs (residues replacing each other) of amino acids, which includes 190 pairs of different amino acid substitutions + 20 pairs of identical substitutions (an amino acid may be replaced back after several replacement cycles during evolution). An example is presented below, the so-called Gonnet matrix:

GH Gonnet, MA Cohen, and SA Benner (1992), Science, Vol 256,1443-1445 1992.
Values rounded to nearest integer

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-1	0	0	0	0	0	-1	-1	0	-1	-2	0	1	1	-4	-2	0		
R	-1	5	0	0	-2	2	0	-1	1	-2	-2	3	-2	-3	-1	0	0	-2	-2	
N	0	0	4	2	-2	1	1	0	1	-3	-3	1	-2	-3	-1	1	0	-4	-1	-2
D	0	0	2	5	-3	1	3	0	0	-4	-4	0	-3	-4	-1	0	0	-5	-3	-3
C	0	-2	-2	-3	12	-2	-3	-2	-1	-1	-2	-3	-1	-1	-3	0	0	-1	0	0
Q	0	2	1	1	-2	3	2	-1	1	-2	-2	2	-1	-3	0	0	0	-3	-2	-2
E	0	0	1	3	-3	2	4	-1	0	-3	-3	1	-2	-4	0	0	0	-4	-3	-2
G	0	-1	0	0	-2	-1	-1	7	-1	-4	-4	-1	-4	-5	-2	0	-1	-4	-4	-3
H	-1	1	1	0	-1	1	0	-1	6	-2	-2	1	-1	0	-1	0	0	-1	2	-2
I	-1	-2	-3	-4	-1	-2	-3	-4	-2	4	3	-2	2	1	-3	-2	-1	-2	-1	3
L	-1	-2	-3	-4	-2	-2	-3	-4	-2	3	4	-2	3	2	-2	-2	-1	-1	0	2
K	0	3	1	0	-3	2	1	-1	1	-2	-2	3	-1	-3	-1	0	0	-4	-2	-2
M	-1	-2	-2	-3	-1	-1	-2	-4	-1	2	3	-1	4	2	-2	-1	-1	-1	0	2
F	-2	-3	-3	-4	-1	-3	-4	-5	0	1	2	-3	2	7	-4	-3	-2	4	5	0
P	0	-1	-1	-1	-3	0	0	-2	-1	-3	-2	-1	-2	-4	8	0	0	-5	-3	-2
S	1	0	1	0	0	0	0	0	0	-2	-2	0	-1	-3	0	2	2	-3	-2	-1
T	1	0	0	0	0	0	0	-1	0	-1	-1	0	-1	-2	0	2	2	-4	-2	0
W	-4	-2	-4	-5	-1	-3	-4	-4	-1	-2	-1	-4	-1	4	-5	-3	-4	14	4	-3
Y	-2	-2	-1	-3	0	-2	-3	-4	2	-1	0	-2	0	5	-3	-2	-2	4	8	-1
V	0	-2	-2	-3	0	-2	-2	-3	-2	3	2	-2	2	0	-2	-1	0	-3	-1	3

Margaret Dayhoff and co-workers, who pioneered the field of protein sequence analysis, databases and bioinformatics, developed the first matrix of this type in the 1970s. Their scoring model was based on observed frequencies of substitutions of each of the 20 amino acids, derived from alignment of closely related sequences. In the resulting **mutation data (or probability) matrix** M_{ij} each element provides an estimate of the probability of an amino acid in column i to be mutated to the amino acid in row j after certain evolutionary time. An evolutionary unit of 100 million years was adapted, resulting in the PAM (percentage accepted mutations / 100 million years) matrix. 1 PAM corresponds to an average amino acid substitution in 1% of all positions. Although 100 PAM does not mean that all the amino acids in the sequence are different, compared to the original sequence, since many of them will be mutated back to their original type. This is logical to assume since preservation of structure and function always have higher priority in the selection process, for this reason there is a limited number of possible replacement at a certain position of a sequence, and the original amino acid is always one of the possible “choices”.

Different versions of the amino acid substitution matrix can be used for different purposes. For example, low PAM (20, 40, 60) may be preferred in database scanning, which uses the so-called **local alignment** algorithms and outputs short alignments of the closest-related sequence segments. The higher the number associated with PAM the longer the evolutionary distance. Thus, high PAM will be suitable for aligning more distant proteins, and if used for database scanning, it will find more distant homologues. It has been shown that at 256 PAM 80 % of all amino acids will be substituted, although to various degrees: 48% of Trp, 41% of Cys and 20% of His would be unchanged, but only 7% of Ser will remain. By other words different amino acids have different propensities for change, presumably due to both structural and functional reasons. For example, as mentioned earlier, tryptophane has a large side chain, and if located within the core of the structure, it would not be easy to replace it by some other amino acid. This may leave a cavity inside the structure, which may destabilize the protein structure as a whole. Cys and His are often involved in some specific functions like protein abstraction (His), metal binding (both), disulfide bridges (Cys), etc, and their replacement will affect the activity of the protein.

Dayhoff matrix was based on a limited set of protein sequences known at that time, and no statistical data could be collected for many of the possible 190 substitutions. This was corrected for in a more recent **PET91 substitution matrix**, essentially an updated Dayhoff matrix (Jones et al., 1992). PET91 was constructed based on a study, which included 2,621 protein families from the SwissProt database (now UniProt, part of the ExPASy server). Meanwhile, other types of substitution matrices were developed, based on slightly different principles. One of the most popular is the **BLOSUM matrix (BLOCKs of Amino Acid SUBstitution Matrix, Henikoff S, Henikoff JG. 1992)**. BLOSUM scores amino acid

replacements based on the frequencies of amino acid substitutions in un-gaped aligned blocks of sequences with a certain percentage sequence identity. This constitutes a major difference between PAM and BLOSUM matrices, since PAM matrices are based on mutations observed in a **global alignment**, which includes highly conserved regions as well as low-conservation regions with gaped alignment. The numbers associated with each matrix (e.g. BLOSUM62, BLOSUM80, etc) refer to the minimum percentage sequence identity of the sequences group within a certain block. Thus, higher numbers correspond to higher sequence identity and shorter evolutionary distance between the proteins. By other words, BLOSUM with high numbers should be used for highly related sequences, while low BLOSUM numbers should be used for distantly related proteins, for example is screening databases.

A number of substitution matrices have also been developed based on the comparison of three-dimensional structures (structure-based alignment). The 3D structure provides information on the position and length of secondary structure elements as well as loop regions, allowing a more precise positioning of gaps. To generate a structure-based sequence alignment it is possible to use a superposition of the 3D structures of the proteins in question (if structures are available for both) or to use the 3D structure of one member of the protein family to guide and correct placement of gaps in a multiple sequence alignment. Many graphics programs include superposition of 3D structures and structure-based alignment of the sequences as an option.

Sequence alignment tutorial 1

Amino acid sequence alignment may be rather simple to run, but may also need some extra attention, for example in cases when the proteins have considerably diverged and there is a large number of insertions and deletions, or in cases of multidomain proteins, especially if not all domains are present in one of the proteins being compared, something which could happen for example during homology modeling. Information from the tertiary structure, like the position of helices, strands and loops, is of course of great help for correct placing of insertions and deletions in the alignment. In this first tutorial we will explore an easy way of making a sequence alignment and will be focusing on using the tools available at Expasy and EBI servers, although there are of course many other servers, which will do exactly the same job. We start with a case of a protein of highly conserved sequence - subunit BchI of the enzyme magnesium chelatase. It is one of three subunits, which are required for this enzyme to catalyze the first committed step in chlorophyll biosynthesis, the insertion of a Mg^{2+} ion into protoporphyrin IX. In the second tutorial we will go through a slightly more complicated case and will first identify the domain of BchD (the second subunit of magnesium chelatase), which is homologous to subunits BchI. We will make then an alignment of the sequences of the two domains to closely examine conservation and differences between the two proteins.

To make the alignment we first need to choose and retrieve the sequences. For this we will use the UniProtKB database within the Expasy group of servers. To start, simply write the name of the protein (BchI) into the UniProt or Expasy search window, and you will be taken to a list of sequences of BchI from different organisms:

UniProtKB results

About UniProtKB 

Filter by¹

 Reviewed (28)
Swiss-Prot
 Unreviewed (1,783)
TrEMBL

Popular organisms

A. thaliana (4)
Rice (1)
RHOCB (1)
RHOS4 (2)
CHLP8 (1)

Other organisms

Search terms

Filter "bchI" as:
gene name (243)
protein name (21)

View by

Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> O50312	BCH1_CHLP8	Magnesium-chelatase 38 kDa subunit	bchI Cpar_0725	Chlorobaculum parvum (strain NCIB 8327) (Chlorobium vibrioforme subsp. thiosulfatophilum (strain DSM 263 / NCIB 8327))	346
<input type="checkbox"/> O30819	BCH1_RHOS4	Magnesium-chelatase 38 kDa subunit	bchI RHOS4_18780, RSP_0273	Rhodobacter sphaeroides (strain ATCC 17023 / 2.4.1 / NCIB 8253 / DSM 158)	334
<input type="checkbox"/> P26239	BCH1_RHOCB	Magnesium-chelatase 38 kDa subunit	bchI RCAP_rcc00677	Rhodobacter capsulatus (strain ATCC BAA-309 / NBRC 16581 / SB1003)	350
<input type="checkbox"/> Q93SW1	BCH1_CHLTE	Magnesium-chelatase 38 kDa subunit	bchI chlI, CT1297	Chlorobium tepidum (strain ATCC 49652 / DSM 12025 / NBRC 103806 / TLS)	392
<input type="checkbox"/> Q9WXA9	BCH1_ACIRU	Magnesium-chelatase 38 kDa subunit	bchI	Acidiphilium rubrum	345
<input type="checkbox"/> J2F3Q4	J2F3Q4_PSEFL	Magnesium chelatase, subunit BchI	bchI PflQ2_3327	Pseudomonas fluorescens Q2-87	333
<input type="checkbox"/> A0A061R6E6	A0A061R6E6_9CHLO	Mg-protoporphyrin IX chelatase	BCHI TSPGSL018_14164	Tetraselmis sp. GSL018	456

The figure is showing just the first few sequences, the actual list contained many more. You may also notice on the left, where it says “Filtered by”, that there are “Reviewed” sequences and “Unreviewed”. It is always better to use the reviewed sequences as much as possible, these have been verified to be what we expect them to be. There are many automatically annotated sequences among the Unreviewed and sometimes they may contain assignment errors.

There we need to choose BCH1_RHOCB (entry P26239), which is subunit BchI from *Rhodobacter capsulatus*. On the page which will open you will find information on the biological function (photosynthesis, magnesium chelatase activity), type of ligands/substrate it binds (ATP), catalytic function (ATP hydrolysis), Protein Data Bank (PDB) entries, if available, links to published works, links to entries related to this particular protein in other databases, and of course the amino acid sequence of the protein. A very useful link is the one to the InterPro database. It provides a plenty of information about the protein, its domain content, biological function, the family to which it belongs, etc. For sequence alignment we first need to retrieve the sequences of BchI from different organisms. Normally the sequence is presented in the following format:

Sequence¹

Sequence status¹: Complete.

P26239-1 [UniParc]  

« Hide

```

      10      20      30      40      50
MTTAVARLQP SASGAKTRPV PFPSAIVGQE DMKLALLTA VDPGIGGLV
      60      70      80      90     100
FGDRGTGKST AVRALAALLP EIEAVEGCPV SSPNVEMIPD WATVLSTNVI
     110     120     130     140     150
RKPTFPVVDLP LGVSEDRVVG ALDIERAISK GEKAFEPGLL ARANRGYLYI
     160     170     180     190     200
DECNLLDHI VDLLLDVAQS GENVVERDGL SIRHPARFVL VSGNPEEGD
     210     220     230     240     250
LRPQLDRFG LSVEVLSPRD VETRVEVIRR RDTYDADPKA FLEWRPKDM
     260     270     280     290     300
DIRNQILEAR ERLPKVEAPN TALYDCAALC IALGSDGLRG ELTLRSARA
     310     320     330     340     350
LAALEGATAV GRDHLKRVAT MALSHRLRDR PLDEAGSTAR VARTVEETLP

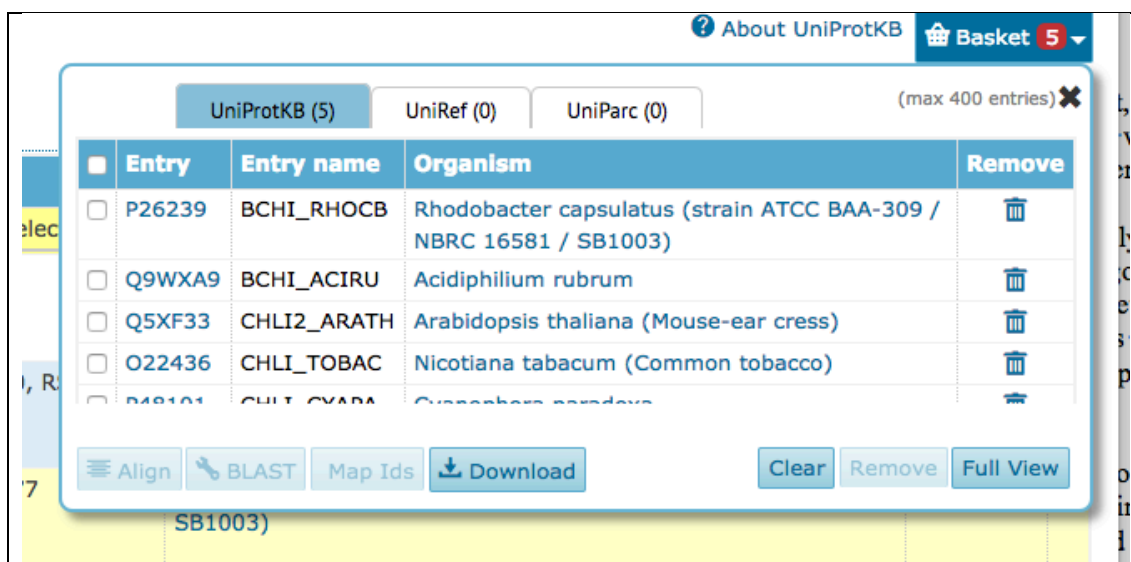
```

Length: 350
Mass (Da): 37,899
Last modified: May 1, 1992 - v1
Checksum: 5CBAA54A1F308568

However, to make an alignment we need to have the sequence in FASTA format, although it will be done automatically when we run the alignment tool at this server. Later we can take the chosen sequences to another server with a better choice of alignment parameters.

To run the alignment we first need to choose some additional sequences. Normally one needs

to spend few minutes and think which sequences to include in the alignment. A good strategy is to include sequence from distantly related organisms, since this will give a better idea on the conservation of the most important amino acids. We can mark the sequences we want to include and click “add to basket” on the top menu. Don’t forget to choose “Reviewed” and to include the protein from *Rhodobacter capsulatus*. The figure below shows the open basket drop-down window. By clicking “Full view” button the drop-down will open as a full web page.



On the full page mark all the sequences and click the “Align” button, which will start the alignment. The results I got look like this:

None

Download Edit and resubmit

Alignment

How to print an alignment in color

Alignment

Tree

Result info

Highlight

Annotation

Chain

Nucleotide binding

Transit peptide

Modified residue

Helix

Disulfide bond

Beta strand

Sequence conflict

Amino acid properties

Similarity

Hydrophobic

Negative

Positive

Aliphatic

Tiny

Aromatic

Charged

Small

Polar

Big

Serine Threonine

Demo

Help video

Entry	Entry name	Organism	Score	Align
P26239	BCHI_RHO�B	Rhodobacter capsulatus (strain ATCC BAA-309 / NBRC 16581 / SB1003)	1	-----MT---TAVARLQPSASGANTREVPFPEAIVGQEDMKLALLTAVDFGIGGVLFV
P48101	CHLI_CYAPA	Cyanothece parvula	1	-----MT---TAVARLQPSASGANTREVPFPEAIVGQEDMKLALLTAVDFGIGGVLFV
Q5XF33	CHLI2_ARATH	Arabidopsis thaliana (Mouse-ear cress)	1	MASLLGRSPSSIL---TCPRISSPSTSSMSHL-----CFGPEKLGRIQFNPKNRSRY
O22436	CHLI_TOBAC	Nicotiana tabacum (Common tobacco)	1	MASLLGTSSAAAAILASTPLSSRSCKPAVSLFSPSSGQGRKFYGGIRVPVKKGRSQY

You may add additional sequences to this alignment (in FASTA format)

I have chosen 4 sequences here. You may also notice that residues involved in ATP binding (nucleotide binding) are marked green, while the rest are coloured in various shades of grey. There are other colouring options available, which we may explore, if required. We may notice that the protein is highly conserved, although the last two sequences have much longer N-terminal part. These two sequence originate from plants, while the first two are from bacteria.

One disadvantage of using this server for alignment is that we cannot change alignment parameters, like the amino acid replacement matrix or gap penalties, discussed earlier. If this is required, one could use the EBI server (European Bioinformatics Institute, <http://www.ebi.ac.uk/Tools/msa/>). We could simply copy and paste the list of the sequences in FASTA format (provided on the page showing the alignment) into the alignment window of the EBI server. On EBI there is also an opportunity to use Jalview, a Java-based application, with which we can color the alignment in different colors, change its appearance in various ways and save a jpg image, e.g. for publication or a presentation. It is possible to use Jalview directly on EBI, however, it is recommended to download and install the application on your own computer. The installed application has much more choices, for example, saving the alignment into an image file for later use. This option is not available in the web-version of the viewer.

Final note: The FASTA format

Many applications require the amino acid sequence to be in FASTA format. The FASTA format includes the amino acid sequence in one-letter code, usually with 60 letters/line. Most important is the sign ">", "larger than", on the first line. Alignment programs like CLUSTALW will use the text after the >-sign on that line as the alignment title for the particular sequence. For convenience, one could leave the name of the protein on that row, which would be useful as a sequence identifier after running the alignment. BchI sequence in FASTA format is shown on the image below:

```
>sp|P26239|BCHI_RHOCB Magnesium-chelatase 38 kDa subunit OS=Rhodobacter capsulatus
MTTAVARLQPSASGAKTRPVFPFSAIVGQEDMKLALLLTAVDPGIGGVLVFGDRGTGKST
AVRALAALLPEIEAVEGCPVSSPNVEMIPDWATVLSTNVIRKPTPVVDLPLGVSEDRVVG
ALDIERAISKGEKAFEPGLLARANRGYLYIDECNLLDHIVDLLLDVAQSGENVVERDGL
SIRHPARFVLVGSGNPEEGDLRPQLLDRFGLSVEVLSPRDVETRVEVIRRRDTYDADPKA
FLEEWRPKMDIRNQILEARERLPKVEAPNTALYDCAALCIALGSDGLRGELTLRSARA
LAALEGATAVGRDHLKRVATMALSHRLRRDPLDEAGSTARVARTVEETLP
```

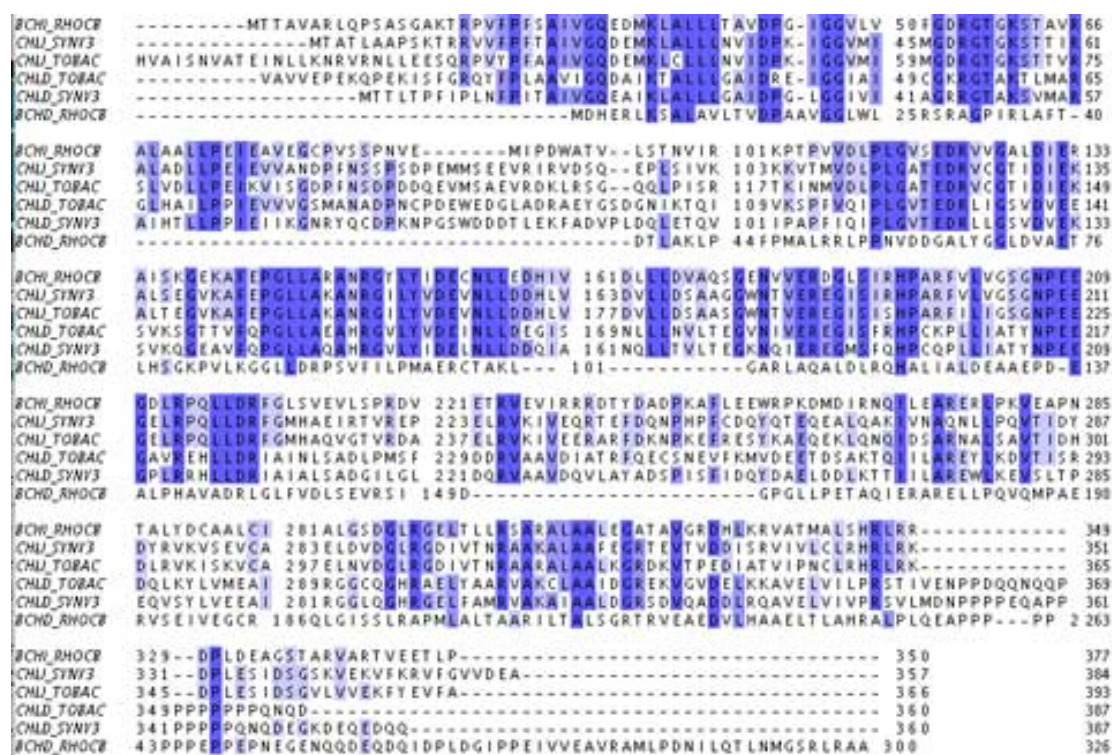
Sequence alignment tutorial 2: BchI-BchD alignment

Now that we have an idea about how to make a simple sequence alignment and how to analyze it, for example by coloring according to percentage identity, coloring only hydrophobic residues, etc, we can look at a more demanding case with some insertions and deletions. This is going to be the second subunit of magnesium chelatase, called BchD, which is almost twice the size of BchI. The task now is to compare the amino acid sequences of the two subunits and find out if they contain homologous parts/domains. We will also learn here how to use secondary structure information in sequence alignment.

For the BchI-BchD sequence alignment it is important to make sure that BchD sequences included in the alignment are really BchD and not something else. Due to automatic annotation procedures used in genomic projects, some proteins, which probably belong to the not so well characterized Ni-chelatase, are annotated as Mg-chelatase. If we would try to make sequence alignment including Ni-chelatase proteins, we will get rather chaotic results with some sequences having only 15% identity to *R. capsulatus* BchD.

To get an idea on the conservation pattern within BchD we can make a separate sequence alignment of BchD. The alignment shown below was run by fetching the sequences from

UniProt and pasting them into the alignment window of MAFFT of the EBI server (don't forget the FASTA format). The “**OUTPUT FORMAT**” was set to “**ClustalW**” and “**ORDER**” to “**Input**”, while “**Gap open penalty**” set to 2.0 to avoid having many small gaps. Image colored using JalView.



We may notice from the alignment above that the N-terminal sequence of these proteins is generally not well conserved - there are several large insertions and deletions in this region. However, the rest of the sequences appears to be well conserved. We could also check the InterPro database to get an idea on domain content of BchD. The analysis will show that the N-terminal part of BchD is an AAA+ domain, which is homologous to BchI (which we need to find out). There is also a von Willebrand type domain at the C-terminus of BchD. This domain has very interesting properties and when we discovered its presence in magnesium chelatase (many years ago) we were very excited. It gave us a lot of clues on the possible functional mechanism of the enzyme. It was discussed in a paper we published on the structure of BchI (Fodje et al, 2000). A later publication describing the complex between subunits BchI and BchD largely confirmed our initial hypothesis (Lundqvist et al, 2010). It is always useful to check the literature to get an idea about the protein before proceeding to sequence alignment and homology modeling.

For the alignment shown below 3 sequences of BchD and 3 of BchI from 3 different organisms were used, and Clustal Omega used for the alignment (don't forget to change the output order to “**input**” instead of the default “align”), **NUMBER of COMBINED ITERATION** and **MAX HMM ITERATIONS** were set to 3. Below I have pasted the N-terminal part of the alignment, which includes the BchI sequence:

33

Introduction to homology modeling

Overview

The term "homology modeling", also called **comparative modeling** or **template-based modeling** (TBM), refers to modeling a protein 3D structure using a known experimental structure of a homologous protein (the template). Structural information is always of great assistance in the study of protein function, dynamics, interactions with ligands and other proteins. The "low-resolution" structure provided by homology modeling contains sufficient information about the spatial arrangement of important residues in the protein and may guide the design of new experiments, for example site-directed mutagenesis. Even within the pharmaceutical industry homology modeling can be valuable in structure-based drug discovery and drug design.

Experimental elucidation of a protein structure may often be delayed by difficulties in obtaining sufficient amount of material (cloning, expression and purification of milligram quantities of the protein) and difficulties associated with crystallization. Even the protein crystallographic part of the project may become a source of problems. In this context, it is not surprising that methods dealing with the prediction of protein structure have gained much interest. Among these methods, the method of homology modeling usually provides the most reliable result. The use of this method is based on the observation that two proteins belonging to the same family (an sharing similar amino acid sequences), will have similar three-dimensional structures. In reality, the degree of conservation of protein three-dimensional structure within a family is much higher than conservation of the sequence.

The steps required in homology modeling are the following:

- template identification;
- amino acid sequence alignment;
- alignment correction;
- backbone generation;
- generation of loops;
- side chain generation & optimization;
- ab initio loop building;
- overall model optimisation;
- model verification. Quality criteria, model quality;

After finding a template it is an absolute requirement to make a multiple sequence alignment, which should include your sequence of course, the sequence of the template and some other sequences of proteins of the same family. This will give an overview of the general features of the protein family, the degree of conservation, the presence and location of consensus sequence motifs, etc. It would also be very desirable to make secondary structure prediction, discussed in the tutorial on sequence alignment. Most importantly, the positions of insertions and deletions should be correct (outside regions of secondary structure), likewise the conserved residues, for example active site residues, should be aligned against each other. When the sequence analysis is done and the alignment is corrected accordingly, we may proceed to the modeling. Modeling software will most probably use its own sequence alignment, which must be checked against your own alignment to make sure that there are now substantial differences. The steps followed by the software include backbone generation, building missing parts (e.g. loops), generation of side chains for residues, optimization of side chain conformations, and energy minimization of the model. The server usually also outputs an assessment of model quality. There are

several servers that may be used for modeling, here we will use the Swiss Model site, which is relatively fast and provides nice model quality assessment. Some other servers, which may use more sophisticated algorithms, can take days (or even weeks!) to return the modeling request. In complicated cases it may be an advantage to use different servers and compare the outputs from them. Of course, the higher the sequence identity between the model and the template the better the expected quality of the model will be.

Like sequence alignment, it is important to keep in mind that depending on the degree of sequence conservation, modeling may be straightforward, but may also be rather challenging, for example, if we need to use 2-3 different templates to model different domains of our protein. The question then will be - how to put these different domains together into one structure? In some cases one could combine modeling with electron microscopy or small-angle X-ray scattering (SAXS) methods, which can provide low-resolution overall shape of the protein in solution. Subsequently, the models of the different domains may be docked into the EM or SAXS densities.

Protein Homology Modeling Using the Swiss Model Server

Modeling with the Swiss Model server

To work with the Swiss Model server (before the start of template identification and modeling), we need to create an account. After providing e-mail address the password will be sent back to the same email. The sequence of the protein to be modeled can be fetched as we did in the sequence alignment tutorial. Then we just paste it into the template identification window and wait for the server to run the Blast search. The Blast search will be run against the sequences of known protein structures from the ExPDB, the SwissModel template library. It is derived from PDB entries, after excluding predicted structures and structures containing only C-alpha atoms. In ExPDB coordinate files containing two or more chains (usually distinguished by chain identifier present in the PDB file after each amino acid name), are split into two or more files, depending on the number of chains (usually denoted A, B, C, etc). For example, PDB entry 1cpc contains two chains, A and B. In ExPDB there will be two entries corresponding to this structure: 1cpcA and 1cpcB. For details on the PDB coordinate file content, please check the related page. Please keep in mind that the Blast run may take some time, all depends on how busy the server is.

As mentioned earlier, it is essential to have an idea on the complexity of the protein homology modeling project before starting the modeling. This can be done by making and analyzing a multiple sequence alignment of your protein with some homologues, including the amino acid sequence (or sequences) of the modeling template identified by the server. As a rule of thumb, a percentage sequence identity above 50% will mean a relatively strait forward modeling project, while anything below that will require careful planning. However, this is just a general rule, it does not mean that careful analysis is not required. There are 5 modeling alternatives available at the Swiss-Model server. The alignment should include a group of homologous sequences from different organisms, including the template (or templates, if more than one) sequences. As discussed in the sequence part, try to choose 3-4 bacterial and an equal amount of eukaryotic sequences. The alignment will show if there are any large insertions and deletions in the protein being analyzed, compared to the template. Insertions (amino acid segments which are not present in the template) mean regions for which the structure is not known and may need to be modeled separately. These also are the regions, which potentially may contain most errors. Usually the server will attempt to model these regions automatically. Loops may be relatively easily modeled, but modeling larger regions is not straightforward.

It is also possible to use the “**Target-Template Alignment**” mode at the server, which lets us

to start with own manually adjusted sequence alignment.

Error sources in homology modeling

The earlier we become aware of possible errors, the better we can eliminate them and handle our modeling project in a proper way. Errors to avoid include the following:

- 1- Incorrect sequence alignment - among the most devastating error in homology modeling.
- 2- Incorrect choice of template - may happen, especially for multi domain proteins.
- 3- Incorrectly built loop regions - loops are usually built automatically by the server. If correct loop conformation is important for the project one could try to do the modeling with different servers and then compared the models from each of them.
- 4- Errors made by the person doing the modeling - this type of errors may include anything and are difficult to predict in advance. Knowledge on the basic principles of protein structure is important for minimizing this type of errors.
- 5- Errors which may be present in the template - difficult to eliminate. A model can hardly be better than the template.

Step by step modeling

In this example, we will make a model for the enzyme magnesium chelatase subunit BchI from Cyanobacteria *Synechocystis* (SWISSPROT entry P51634).

The results of the Blast search are shown below:

Template Results

Template Results							
<div> <div>Templates</div> <div>Sequence Similarity</div> <div>Alignment of Selected Templates</div> <div>More ▾</div> </div>							
◆ Name ◆	Title	◆ Coverage ◆	◆ Identity ◆	◆ Method ◆	◆ Oligo State ◆	Ligands	
<input checked="" type="checkbox"/> 2x31.1.L	MAGNESIUM-CHELATASE 38 KDA SUBUNIT		52.68	EM, 7.5Å	hetero-oligomer	None	
<input type="checkbox"/> 2x31.1.K	MAGNESIUM-CHELATASE 38 KDA SUBUNIT		52.68	EM, 7.5Å	hetero-oligomer	None	
<input type="checkbox"/> 1g8p.1.A	MAGNESIUM-CHELATASE 38 KDA SUBUNIT		52.68	X-ray, 2.1Å	monomer	None	
<input type="checkbox"/> 1g8p.1.A	MAGNESIUM-CHELATASE 38 KDA SUBUNIT		52.47	X-ray, 2.1Å	monomer	None	
<input type="checkbox"/> 2x31.1.K	MAGNESIUM-CHELATASE 38 KDA SUBUNIT		52.47	EM, 7.5Å	hetero-oligomer	None	
<input type="checkbox"/> 2x31.1.L	MAGNESIUM-CHELATASE 38 KDA SUBUNIT		52.47	EM, 7.5Å	hetero-oligomer	None	
<input type="checkbox"/> 3hte.1.D	ATP-dependent Clp protease ATP-binding subunit clpX		16.93	X-ray, 4.0Å	homo-hexamer	None	
<input type="checkbox"/> 3hte.1.B	ATP-dependent Clp protease ATP-binding subunit clpX		16.93	X-ray, 4.0Å	homo-hexamer	None	
<input type="checkbox"/> 3hte.1.A	ATP-dependent Clp protease ATP-binding subunit clpX		16.93	X-ray, 4.0Å	homo-hexamer	None	
<input type="checkbox"/> 3hte.1.C	ATP-dependent Clp protease ATP-binding subunit clpX		16.93	X-ray, 4.0Å	homo-hexamer	None	

The best choice in the list is *R. capsulatus* BchI (PDB ID 1g8p). The other proteins in the list are of very low resolution (7.5 Å) and originate from electron microscopic (EM) modeling. They are essentially the same protein (1g8p), which has been slightly modified to fit the EM model. Longer in the list there are other structures. The sequence identity with our protein is around 20%, and none of them is a magnesium chelatase. However, they are interesting since they are members of the family of AAA ATPases one could have a closer look at them at some later stage, they may shed additional light at the mechanism of magnesium chelatase. These proteins are involved in a large number of biochemical processes in organisms.

By clicking the arrow in the right, you will get the alignment of the target sequence with the template:

2x31.1.L MAGNESIUM-CHELATASE 38 KDA SUBUNIT 52.68 EM, 7.5Å hetero-oligomer None

Method ELECTRON MICROSCOPY 7.50 Å

Found By BLAST

GMQE 0.73

Seq Similarity 0.44

Oligo State Hetero-oligomer

Target Prediction Oligomeric state prediction is currently only available for homo-oligomers.

Build Model



Target MTATLAAPSKTRRVFFFTAIVGQDEMKLALLLNVIDPKIGGVMMGDRGTGKSTTIRALADLLP 65
 2x31.1.L ----- A R V P P F A I V G Q M K L A L L L A V D P I G G V M G D R G T G K S T A V R A L A A L L P 70
 Target EIEVVANDPFNSSPSPDEMMSEEVRI R V D S Q E P L S I V K K K V T M V D L P L G A T E D R V C G T I D I E K 128
 2x31.1.L E I E A V M G C P V S P N ----- V E M I P D W A T V L S T N V I E K T P V V D L P L G V S E D R V V G A D I E R 126
 Target ALSEG VK A F E P G L L A K A N R G I L Y V D E V N L L D D H L V D V L L D S A A G G W N T V E R E G I S I R H P A R F V L V 193
 2x31.1.L A I S R G E K A F E P G L L A R A N R G I L Y D E C N L L D H I V D L L D V A Q S G E N V V E R D G L S I R H P A R F V L V 191
 Target G S G N P E E G E L R P Q L L D R F G M H A E I R T V R E P E L R V K I V E Q R T E F D Q N P H P F C D Q Y Q T E Q E A L Q A K I 258
 2x31.1.L G S G N P E E G E L R P Q L L D R F G L S V E V L S P R D V E T R V E V I R R R D T Y D A P K A F L E E W R P K D M D I R N Q I 256
 Target V N A Q N L L P Q V T I D Y D Y R V K V S E V C A E L D V D G L R G D I V T N R A A K A L A A F E G R T E V T V D D I S R V I V L 323
 2x31.1.L L E A R E L P K V S A P N I A L Y D C A A L C I A L G S D G L R G E L L L R A R A L A A L E G A T A V R D H L K R V A T H I 321
 Target C L R H R L R K D P L E S I D S G S K V E K V F K R V F G V V D E A 357
 2x31.1.L A L S H R L R R D P L D E A G S T A R V A R I V E T I ----- 349

We may choose to build a model, however, we need first to check the sequence alignment to ensure that everything is correct. The CHLI_SYNY3 sequence was actually included in the sequence alignment exercise we made earlier and if we compare the alignments we may see that the largest gap is slightly shifted to the left here. This is probably ok and we can proceed with the modeling. We can always come back to this if we are not satisfied with the model for some reason.

The modeling results output is shown on the following image:

Method X-RAY DIFFRACTION 2.10 Å

Found By HHblits


GMQE 0.66

Seq Similarity 0.44

Oligo State Monomer

Target Prediction Oligomeric state prediction is currently only available for homo-oligomers.

Build Model



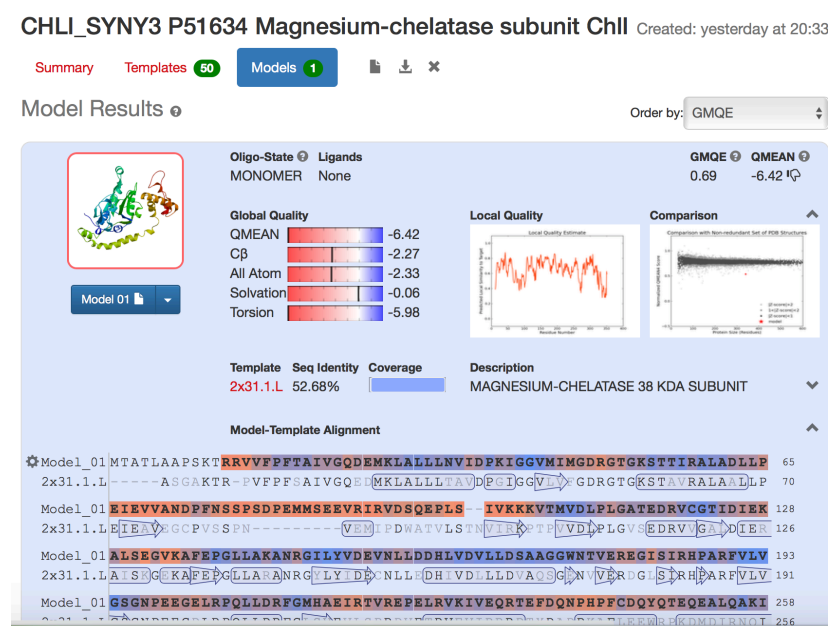
Target MTATLAAPSKTRRVFFFTAIVGQDEMKLALLLNVIDPKIGGVMMGDRGTGKSTTIRALADLLP 65
 1g8p.1.A ----- A R V P P F A I V G Q M K L A L L L A V D P I G G V M G D R G T G K S T A V R A L A A L L P 70
 Target EIEVVANDPFNSSPSPDEMMSEEVRI R V D S Q E P L S I V K K K V T M V D L P L G A T E D R V C G T I D I E K A I 130
 1g8p.1.A E I E A V M G C P V S P N ----- V E M I P D W A T V L S T N V I E K T P V V D L P L G V S E D R V V G A D I E R A I 128
 Target SEG VK A F E P G L L A K A N R G I L Y V D E V N L L D D H L V D V L L D S A A G G W N T V E R E G I S I R H P A R F V L V G S 195
 1g8p.1.A S E G K A F E P G L L A R A N R G I L Y D E C N L L D H I V D L L D V A Q S G E N V V E R D G L S I R H P A R F V L V G S 193
 Target G N P E E G E L R P Q L L D R F G M H A E I R T V R E P E L R V K I V E Q R T E F D Q N P H P F C D Q Y Q T E Q E A L Q A K I V N 260
 1g8p.1.A G N P E E G E L R P Q L L D R F G L S V E V L S P R D V E T R V E V I R R R D T Y D A P K A F L E E W R P K D M D I R N Q I L E 258
 Target A Q N L L P Q V T I D Y D Y R V K V S E V C A E L D V D G L R G D I V T N R A A K A L A A F E G R T E V T V D D I S R V I V I C I 325
 1g8p.1.A A R E L P K V S A P N I A L Y D C A A L C I A L G S D G L R G E L L L R A R A L A A L E G A T A V R D H L K R V A T H I 323
 Target R H R L R K D P L E S I D S G S K V E K V F K R V F G V V D E A 357
 1g8p.1.A R H R L R D P L D E A G S ----- 337

<input type="checkbox"/>	2x31.1.K	MAGNESIUM-CHELATASE 38 KDA SUBUNIT	52.47	EM, 7.5Å	hetero-oligomer	None	▼
<input type="checkbox"/>	2x31.1.L	MAGNESIUM-CHELATASE 38 KDA SUBUNIT	52.47	EM, 7.5Å	hetero-oligomer	None	▼
<input type="checkbox"/>	3hte.1.D	ATP-dependent Cip protease ATP-binding subunit cipX	16.93	X-ray, 4.0Å	homo-hexamer	None	▼
<input type="checkbox"/>	3hte.1.B	ATP-dependent Cip protease ATP-binding subunit cipX	16.93	X-ray, 4.0Å	homo-hexamer	None	▼
<input type="checkbox"/>	3hte.1.A	ATP-dependent Cip protease ATP-binding subunit cipX	16.93	X-ray, 4.0Å	homo-hexamer	None	▼

Here we should start analyzing the quality of the model. You may notice that the QMEAN

value is satisfactory, and if we check the local quality plot we will immediately notice the source of a potential problem (low similarity values) is located in at the beginning of domain 1, which is essentially the position of the insertion shown in the alignment above. The graphics model on the right (not visible in the images above) shows this region in red color. Clicking on the question mark close to QMEAN will open a page with explanation of its meaning. In the model quality part we will have a closer look at these problems.

What would have happened if we would choose the 7.5 Å resolution structure for modeling? The answer is in the image below, which shows considerably lower model quality (high QMEAN) with problems both at the beginning and end of the structure, as seen on the local quality plot:



Quality assessment of a homology model

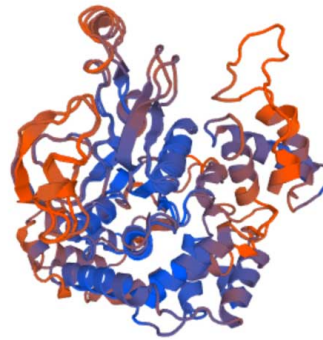
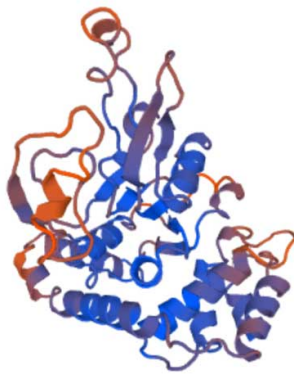
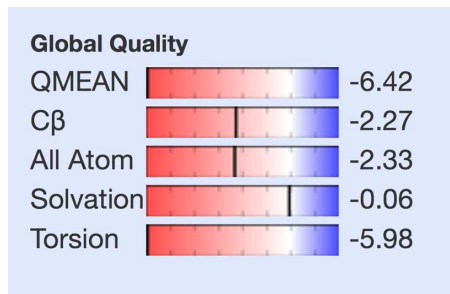
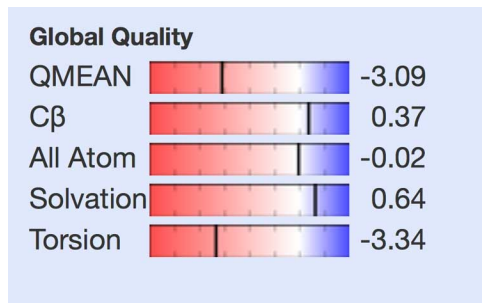
Let us have a look at the output we get from the Swiss Model server for the Bchl modeling project.

What does all that tell us?

For assessing the quality of the homology model the server provides several scores. Clicking the question mark will tell us that QMEAN4 scoring function (the original paper Benkert et al. 2008, and a more recent paper Benkert et al., 2012) is a linear combination of four structural descriptors:

- The local geometry is analyzed by a torsion angle potential over three consecutive amino acids.
- Two distance-dependent interaction potentials are used to assess long-range interactions: First, at a residue-level it is based on C-beta atoms only, at the second level an all-atom potential is used.
- A solvation energy is calculated to investigate the burial status (accessibility to water) of the residues.

In the paper about QMEAN it is stated that "QMEAN shows a statistically significant improvement over nearly all quality measures describing the ability of the scoring function to identify the native structure and to discriminate good from bad models". In the image below we can see the difference between the "good" model (left) and "bad model" (right) we made in the previous page:



The figure above is colored according to error values - low-error regions blue and high-error regions red. We may recognize the largest red-color region as the one where the extra 7 amino acid insertion in the sequence is located. It was built by the server according to the alignment. Since we do not have any experimental data to improve the structure, we could, for example, try to find another server, specialized in building this type of models. There are different possibilities, this region could be rearranged into a β -hairpin or it could also include a short α -helix - a secondary structure prediction may give some indication on that.

It could also be tested experimentally, if we had both *R. capsulatus* Bchl and SyncChlI proteins expressed and purified, we could do some CD spectroscopic measurements to compare secondary structure content of the two proteins. For example higher percentage β -structure in SyncChlI would indicate that this region may be a β -hairpin. One could also try to run some molecular dynamics simulations on the protein and see if this region would converge to some other structure. The resulting QMEQN can always be checked at the QMEAN server.

Of course the best would be to crystallize the protein and determine its structure - it will immediately reveal the differences

Further assessment of homology model quality

The tools available in SwissPDB Viewer may also be used for quick model evaluation. The program uses mean force potential calculations of the energy of each residue in the model and displays the results in the form of a graph. To do this we need to load the optimized homology model file into SPDBV. You may open the homology model file and the experimental 1g8p to see if there are any differences. The easiest will be to display the structures as C α -alpha trace and color them differently. You will immediately notice the two regions where the structures differ from each other.

For energy calculations make sure the current layer is the **model**, and click on the little white arrow located at the right of the help question mark of the Align Window. The window expands, and displays a curve showing the energy of each residue (interactions with surrounding atoms). If there are no bad contacts, the energy is around or below zero, whereas bad contacts will have high energy above the zero line (red regions). You may also chose "Color: Force filed", which will color the model and show regions with high energy. However, the energy analysis provided by the modeling server is probably more comprehensive.

At the early stages of homology modeling you can also evaluate how good your model is by using the "select aa making clashes" items of the "Select" menu. This will allow you to quickly focus on potentially problematic regions (holding the option key while you select these will not only select aa but also draw the clashes in pink on the screen). You can then choose the "Fix Selected Side chains (quick and dirty)" item of the "Tools" menu, which will browse the rotamer library to choose the best rotamer (the same commands are used if you want to replace an amino acid by another). By repeating the "Select aa making clashes" process, you should see that far less amino-acids are making problems. If not, this is probably a good clue that your threading (by other words the sequence alignment) is incorrect.

Important Note: Fixing side chains is just for you to evaluate the preliminary model prior to submitting it to the server. It will have little influence on model building and the quality of the final model, as the server reconstructs side chains during that process.

Other criteria for the quality of the homology model include model geometry and particularly the Ramachandran plot. The Ramachandran plot may be checked in DeepView, you need to go to the Wind menu of the program and choose "Ramachandran", then in the "Select" menu choose "ALL". This will display all the torsion angles of the model in the Ramachandran plot. Pointing at any point in the plot will show the residue name. This way we may be able to check if there are any residues with bad torsion angles. A more comprehensive way of checking the geometry of the model is to use one of the dedicated servers, for example the JCSG Protein structure validation server (Joint Center for Structural Genomics) and submit your model for evaluation using programs like Procheck:

PROTEIN STRUCTURE VALIDATION

Instructions:

- Check the programs you want to run
- Provide necessary files
- Provide your email address
- Submit and wait for results

☐ PROCHECK [v.3.5.4](#)

☐ SFCHECK [v.6.0.2](#)

ingen fil vald Provide [CIF format](#) structure factor file

Or

ingen fil vald Provide mtz file

For mtz file, please Provide the labels

☐ WHATCHECK [v.19991018-1516](#)

☐ ERRAT

☐ DDQ [v.2.0](#) **NOTE:** To run DDQ, the PDB file must contain SCALE data.

ingen fil vald Provide file containing the positive difference peaks generated by [CNS](#)

ingen fil vald Provide file containing the negative difference peaks generated by [CNS](#)

☐ PROVE [v.2.5.1](#)

☐ WASP

ingen fil vald Provide coordinates file in PDB format

Provide your email address

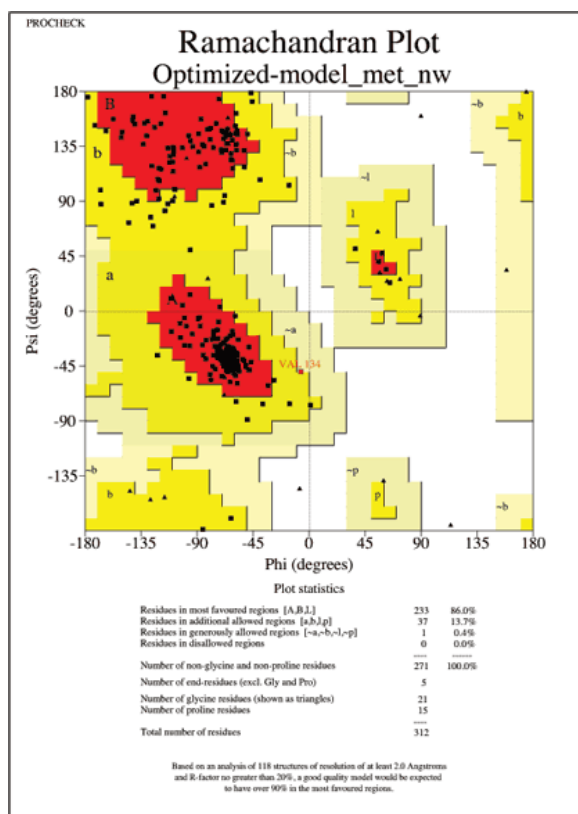
After submitting you will get an e-mail with a link to the structure validation. My output looked like this:

```
-----PROCHECK.....SUMMARY-----
...I...Optimized-model_met_nw.pdb...2.0...312...residues...
+I...Ramachandran Plot:.....86.0%...core.....13.7%.....allow.....0.4%.....gener.....0.0%.....disall
*I...All Ramachandrans:..... 17...labelled.....residues.....(out.....of.....307)
+I...Chi1-chi2 plots:..... 4...labelled.....residues.....(out.....of.....186)
...I...Main-chain params:..... 6...better.....0.....inside.....0.....worse
...I...Side-chain params:..... 5...better.....0.....inside.....0.....worse
*I...Residue properties:..... Max.deviation:...11.7.....Bad.....contacts:.....11
*I.....Bond..len/angle:.....7.6.....Morris...et...al...class:.....1...1...2
...I...Main chain bond lengths:..... 99.7%...within.....limits.....0.3%.....highlighted
...I...Main chain bond angles:..... 97.3%...within.....limits.....2.7%.....highlighted
*I...Planar groups:.....83.6%.....within.....limits.....16.4%.....highlighted
```

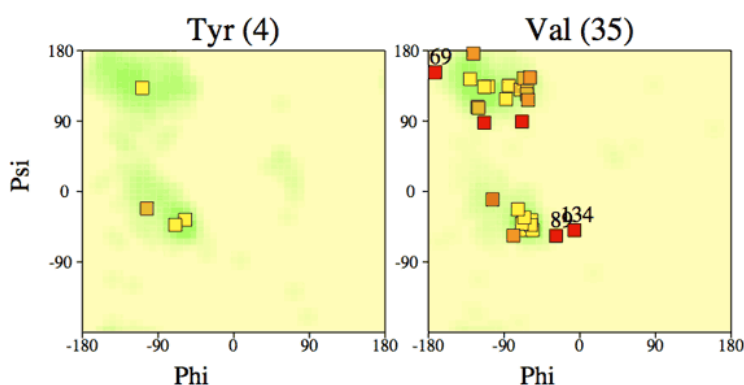
+ May be worth investigating further. * Worth investigating further.

[Detailed proccheck output](#)

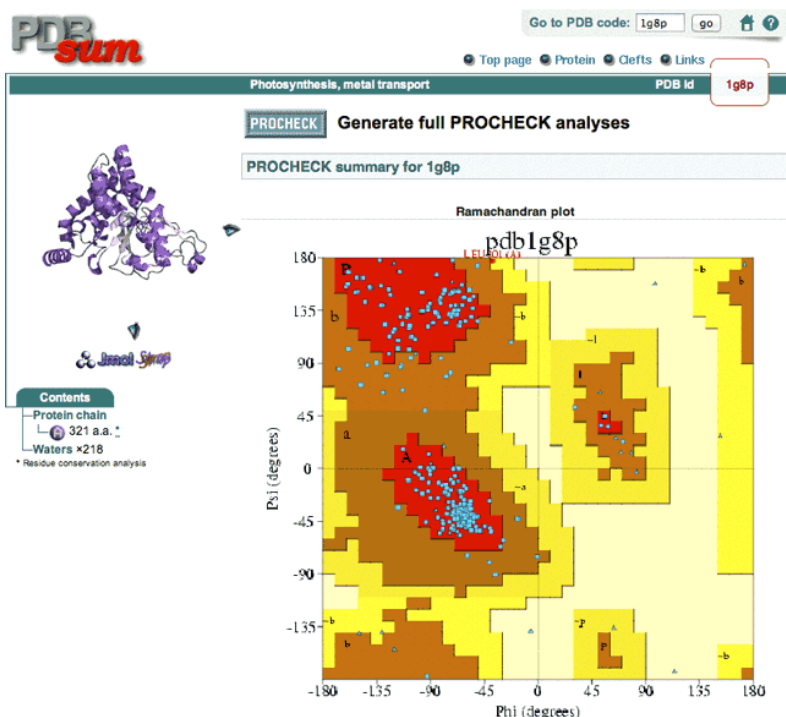
By clicking on any of the links above one would get a detailed description of the corresponding parameter. For example, the Ramachandran plot:



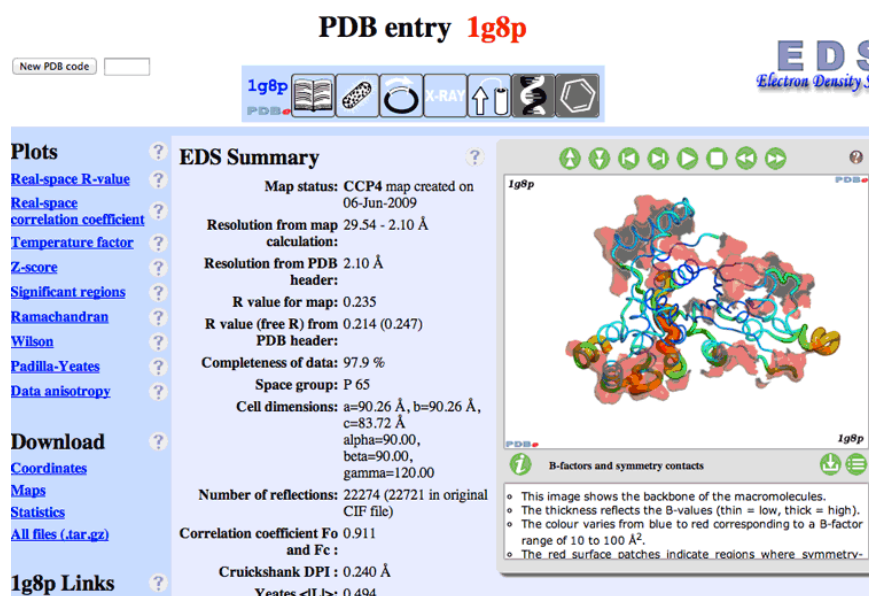
And also detailed analysis showing the torsion angles for all amino acids in the protein (one aa type a time, total of 20), like here for Val and Tyr:



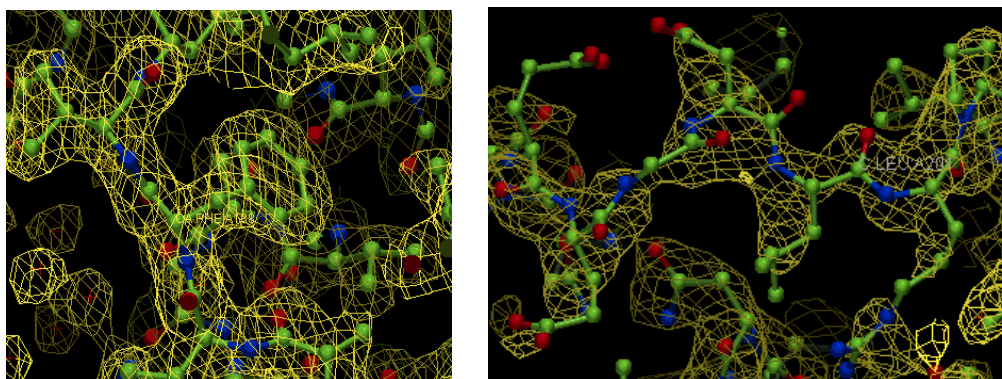
The analysis indicates that there are 17 residues with bad torsion angles. This is not surprising, we got an indication that this was the case when we looked at the energy in the previous page. One probably needs to go back and check these residues carefully and try to understand the problem. Another possibility is to have a look at the Procheck output for the modeling template, in this case 1g8p. To do that we don't need to submit the coordinates to the verification server, we can just go to PDBsum, find 1g8p and click the Ramachandran plot symbol on the right, and on the page which will appear click "Procheck" next to generate Procheck Analysis:



The analysis report will show that essentially the same residues in the 1g8p structure have problems with their Ramachandran angles. By other words, the homology model has just inherited the problems of the template! We could also be more curious and check the electron density for the experimental structure to see if it has any problems within the regions where we get bad torsion angles. To do that we need to use the electron-density server, EDS. While there, enter the PDB code, which brings us to the following page:



Down in this page we can start the Astex viewer, which will display the electron density of the molecule. We may easily center on the residues we are interested in by clicking on the sequence below the graphics window (not shown here). There are several options in the program which may modify the view of the model and the electron density. Below on the left one of the regions which has good quality electron density is shown and on the right one of the regions, which had bad torsion angles, and apparently weak electron density is shown:



The weak density has contributed to bad geometry in this region of the structure. This example shows how important the quality of the electron density is for model quality. It also shows that problems present in the template are imported to the homology model.