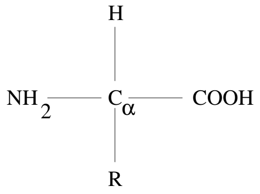
7 Protein secondary structure

Sources for this chapter, which are all recommended reading:

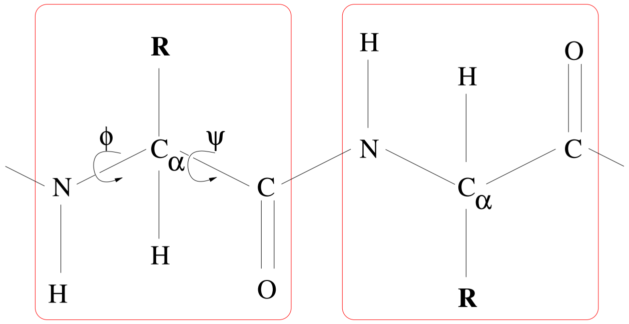
- V.V. Solovyev and I.N. Shindyalov. Properties and prediction of protein secondary structure. In *Current Topics in Computational Molecular Biology*, T. Jiang, Y. Xu and M.Q. Zhang (editors), MIT press, chapter 15, pages 366-401, 2002.
- D.W. Mount. *Bioinformatics: Sequences and Genome analysis*, Cold Spring Harbor Press, Chapter 9: Protein classification and structure prediction. pages 381-478, 2001.

7.1 Proteins

A *protein* is a chain of amino acids joined by peptide bonds. It is produced by a ribosome that moves along an mRNA and adds amino acids according to the codons that it observes.

The structure of an amino acid: 

Here are two amino acids within a polypeptide chain:



The R group is different for each of the twenty amino acids. The R groups of a protein are called its *side chains*. Neighboring amino acids are joined by a peptide bond between the C=O and NH groups. A chain of repeated N-C_α-C's make up the *backbone* of the protein.

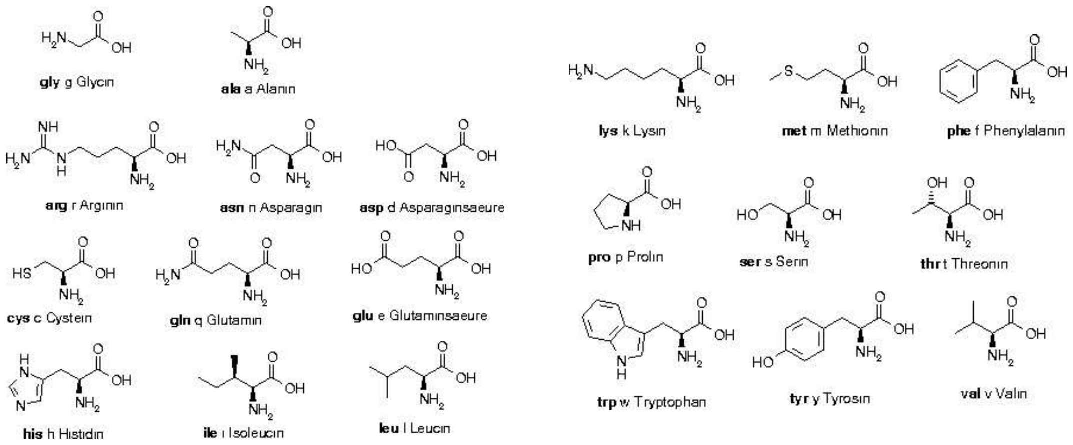
In such a polypeptide chain, each amino acid has two rotational degrees of freedom:

- the rotational angle ϕ of the bond between N and C_α, and
- the rotational angle ψ of the bond between C_α and C.

Both bonds are free to rotate, subject to spatial constraints posed by adjacent R groups. The third angle Ω of the peptide bond between the C=O and NH groups is nearly always 180°.

A protein starts with a free NH group (the *N-terminus*) and ends with a free COOH group (the *C-terminus*).

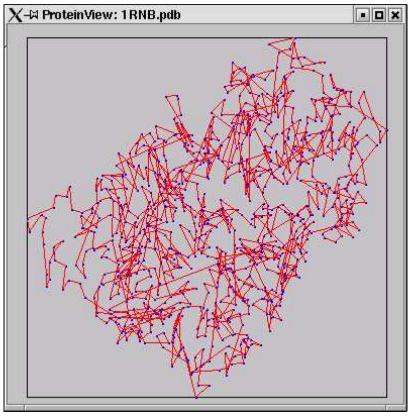
The twenty common amino acids:



(Figure from: `http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids_en.html`)

7.2 Visualization of proteins

The aim of assignment sheet 12 is that you familiarize yourselves with PDB files and the data associated with three-dimensional protein structures. You are requested to write a simple viewing program that displays the three-dimension structure like this:



(Note that of the three given pdb files, only `2BOP.pdb` contains secondary structure information (`HELIX`, `SHEET` and `TURN` lines).)

7.3 Hierarchy of protein structure

K.U. Linderstrom-Lang (Linderstrom-Lang & Schnellman 1959) proposed to distinguish four levels or protein structure:

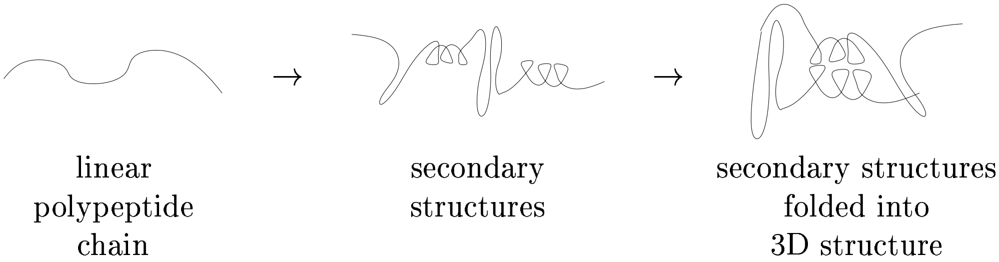
- The *primary structure* is the chemical structure of the polypeptide chain(s) in a given protein, i.e. its sequence of amino acid residues that are linked by peptide bonds.

- The *secondary structure* is folding of the molecule that arises by linking the C=O and NH groups of the backbone together by means of hydrogen bonds.

- The *tertiary structure* is the three dimension structure of the molecule consisting of secondary structures linked by “looser segments” of the polypeptide chain stabilized (primarily) by side-chain interactions.

- The *quaternary structure* is the aggregation of separate polypeptide chains into the functional protein.

Pathway for folding a linear chain of amino acids into a three-dimensional protein structure:



The tertiary structure of proteins is of great interest, as the shape of a protein determines much, if not all, of its function.

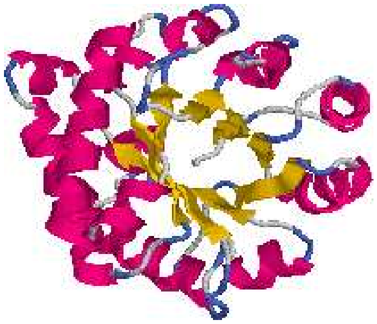
At present, the experimental determination of protein structure via x-ray crystallography is difficult and time-consuming. Hence, we would like to be able to determine the structure of a protein from its sequence. Determining the secondary structure is an important first step.

7.4 The “Holy Grail” of bioinformatics

The *holy grail* of bioinformatics: develop and algorithm that can reliably predict the structure (and thus function) of a protein from its amino-acid sequence!

```
...
IIFIATTNLLGLLPHSFTPTTQLSMNLAMAIPLWA
GAVILAHFLPQGTPTPLIPMLVIIETISLLIQPAL
AVRLTANITAGHLLMGSATLAMTLIIFTILILLTI
LEIAVALIQAYVFTLLVSLYLHDNTPQLNTTVWPT
MITPMLLTLFLITQLKMLPWEPKWADRWLFSTNHK
DIGTLYLLFGAWAGVLGTALSLLIRAELGQPGNLL
GNDHIYNVIVTAHAFVMIFFMVMPIMIGGFGNWLV
PLMIGAPDMAFPRMNNMSFWLLPPSLLLLLASAMV
...
```

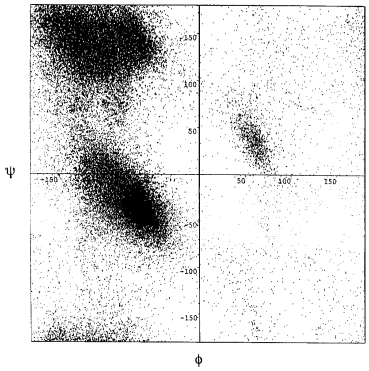
$\xrightarrow{?}$



This is not easy...

7.5 Secondary structure of proteins

Regular features of the main chain of a protein give rise to a secondary structure. The two most common regular structures are called α *helix* and β *sheet* (L. Pauling 1951), corresponding to specific choices of the ϕ and ψ angles along the chain. This can be seen in a *Ramachandran plot* of observed pairs of angles in a collection of known protein structures:



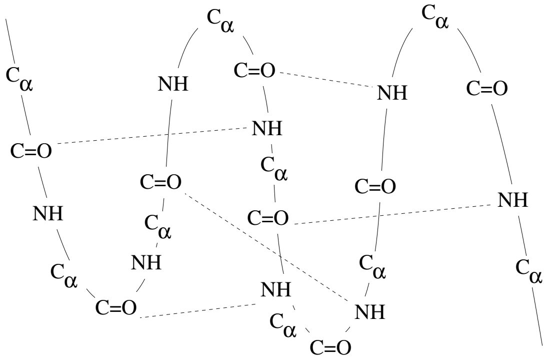
The pairs near $\phi = -60°$ and $\psi = -60°$ correspond to α helices. The pairs near $(-90°, 120°)$ correspond to β strands.

7.6 α helices and β sheets

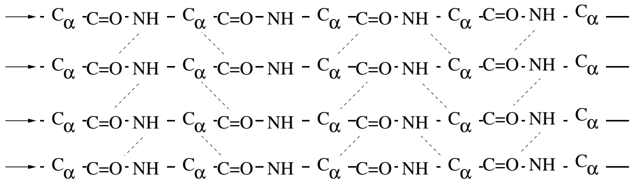
Helices arise when hydrogen bonds occur between (the C=O group of) the amino acid at position i and (the NH group of) the amino acid at position $i + k$ (with $k = 3, 4$ or 5), for a run of consecutive values of i.

Most often, $k = 4$ or 5 and the resulting structure is called an α *helix*, whereas $k = 3$ gives rise to a 3_{10} *helix*.

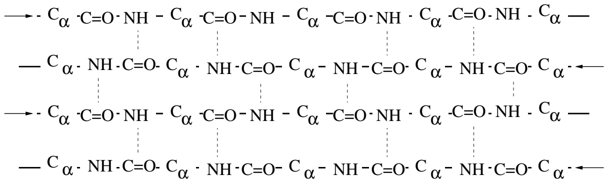
Here is the bonding pattern of an α helix:

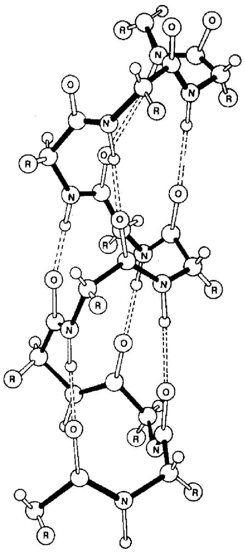


So-called β *sheets* are formed by H bonds between a run of 5-10 consecutive amino acids in one portion of the chain and a another 5-10 consecutive amino acids further down the chain. There are two possible configurations. In a parallel β sheet, all chains run in the same direction:

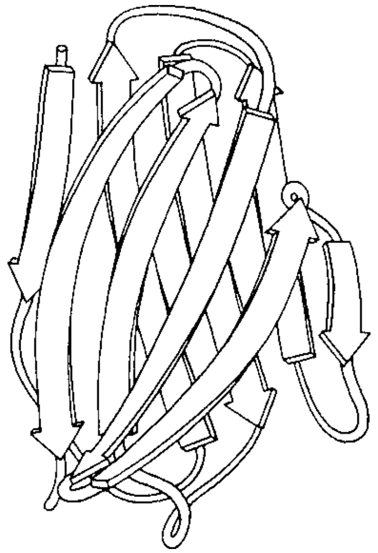

→ C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α —
→ C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α —
→ C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α —
→ C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α —

In an anti-parallel sheet, chains run in alternating directions:


→ C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α —
— C_α- NH-C=O- C_α - NH-C=O- C_α- NH-C=O- C_α- NH-C=O- C_α ←
→ C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α -C=O-NH - C_α —
— C_α- NH-C=O- C_α - NH-C=O- C_α- NH-C=O- C_α- NH-C=O- C_α ←



α helix
3.6 residues per turn



anti-parallel β sheet
(V2 domain of an immunoglobulin)

(Images from: `http://www.tau.ac.il/~becker/course/secondary.html`)

7.7 Loops and coils

Loops or *turns* are regions of a protein chain that lie between α helices and β sheets. The lengths and three-dimensional structure of loops can vary. Hairpin loops that constitute a complete turn joining two anti-parallel β-strands may be as short as two amino acids. They lie on the surface of the structure.

Because they are not spatially constrained and do not affect the arrangement of the secondary structure elements, more substitutions, insertions and deletions can occur than inside the *core* of the protein that is made up of the α helices and β sheets.

A region of secondary structure that is not a helix, a sheet, or recognizable turn is called a *coil*.

7.8 Secondary structure of proteins

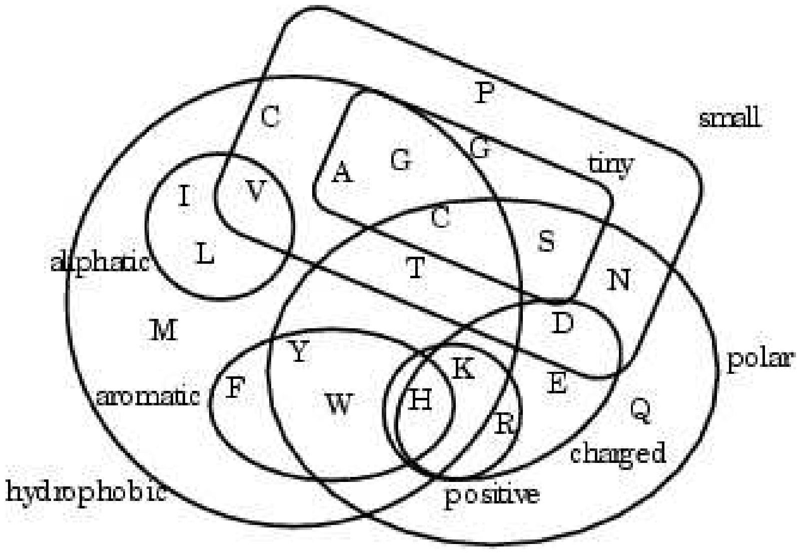
About 35% of residues in proteins lie in α helices. The amino acids Ala, Glu, Leu and Met are often found in α helices, whereas Pro, Gly, Ser, Thr and Val occur rarely in them.

Most α helices are immersed in the protein interior from one side and form a exterior protein surface from the other side. Non polar residues are usually located on one side of an α helix (forming a hydrophobic cluster) and polar and charged residues are on the other side.

About 36% of all residues of proteins are contained in β sheets. The C_α atoms and side chains alternate above and below the sheet in a pleated structure. If one side of a β sheet is exposed to solvent, then this results in a characteristic interchange of hydrophobic and polar amino acids. Parallel β sheets are usually found in the interior of a protein, often protected from solvents by α helices on both sides. They usually incorperate at least 5 strands.

Amino acids Val, Ile, Try and Thr prefer β sheets, whereas Glu, Gln, Lys, Asp, Pro and Cys are seldomly found in them.

7.9 Classification of amino acids

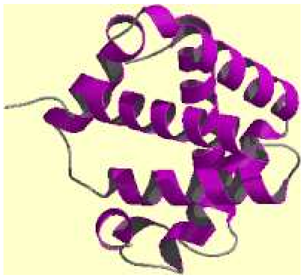


7.10 Classification of protein structure

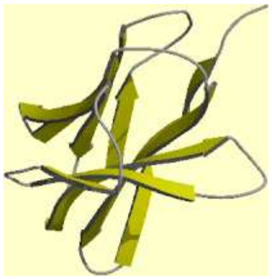
D.W. Mount describes six principal classes of protein structures, four from Levitt and Chothia (1976), and two additional ones taken from the SCOP database (Murzin et al., 1995):

(1) A member of class α consists of a bundle of α helices connected by loops on the surface of the proteins.

(2) A member of class β consists of β sheets, usually two sheets in close contact forming a sandwich. Examples are enzymes, transport proteins and antibodies.



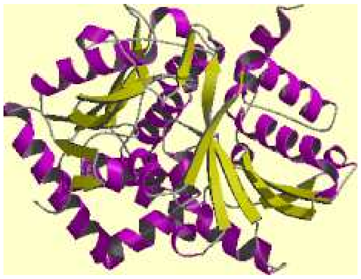
(1) Hemoglobin (3hhb)



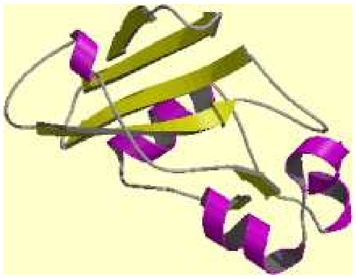
(2) T-cell receptor CD8 (1cd8)

(3) A member of class α/β consists mainly of β sheets with intervening α helices. This class contains many metabolic enzymes.

(4) A member of class $\alpha + \beta$ consists of segregated α helices and β sheets.



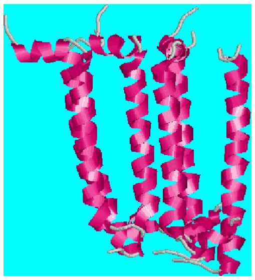
(3) Tryptophan synthase β subunit (2tsy)



(4) G-specific endonuclease (1rnb)

(5) This class consists of all multi-domain (α and β) proteins with domains from more than one of the above four classes.

(6) Membrane and cell-surface proteins and peptides excluding proteins of the immune system.



(6) Integral membrane light-harvesting complex (1kzu)

Figures from `http://www.biochem.ucl.ac.uk/bsm/cath_new/`.

The three letter codes are PDB codes.

7.11 Detecting the secondary structure of a known 3D structure

Given the positions of the main chain atoms of a protein. For each amino acid in the protein, we want to determine whether it belongs to an α helix or a β sheet, or neither. The *DSSP (definition of secondary structure of proteins)* algorithm (Kabsch, Sander, 1983) proceeds as follows:

First determine which C=O and NH groups in the main chain are joined by hydrogen bonds. This decision is based on an electrostatic model using the following energy calculation:

$$E = q_1 q_2 \left(\frac{1}{r(ON)} + \frac{1}{r(CH)} - \frac{1}{r(OH)} - \frac{1}{r(CN)} \right) \cdot f \cdot N_A,$$

with $q_1 = 0.42e$ and $q_2 = 0.20e$, where $e = 4.80325 \times 10^{-10} gr^{\frac{1}{2}} cm^{\frac{3}{2}} s^{-1}$ is the unit electron charge, $r(AB)$ is the inter-atomic distance from atom A in the first amino acid and atom B in the second in Angstroms, $f = 332$ is a constant called the *dimensionality factor*, and E is the energy in kcal/mol.

7.12 Detecting α helices

Recall that Avogadro's constant $N_A = 6.02214199 \times 10^{23} \text{mol}^{-1}$ equals the number of molecules in one mole of substance.

For any pair of amino acids, for which $E < -0.5$ kcal/mol, an H-bond is assigned.

Any H-bond detected in this way is called

- an *n-turn*, if it connects the O=C group of amino acid i to the NH group of amino acid $i + k$, where $k = 3, 4$ or 5, and
- a *bridge*, if it connects residues that are not near to each other.

An α helix is identified as a consecutive sequence of (at least two) n-turns, Any two helices that are offset by two or three residues are concatenated into a single helix.

7.13 Detecting β sheets

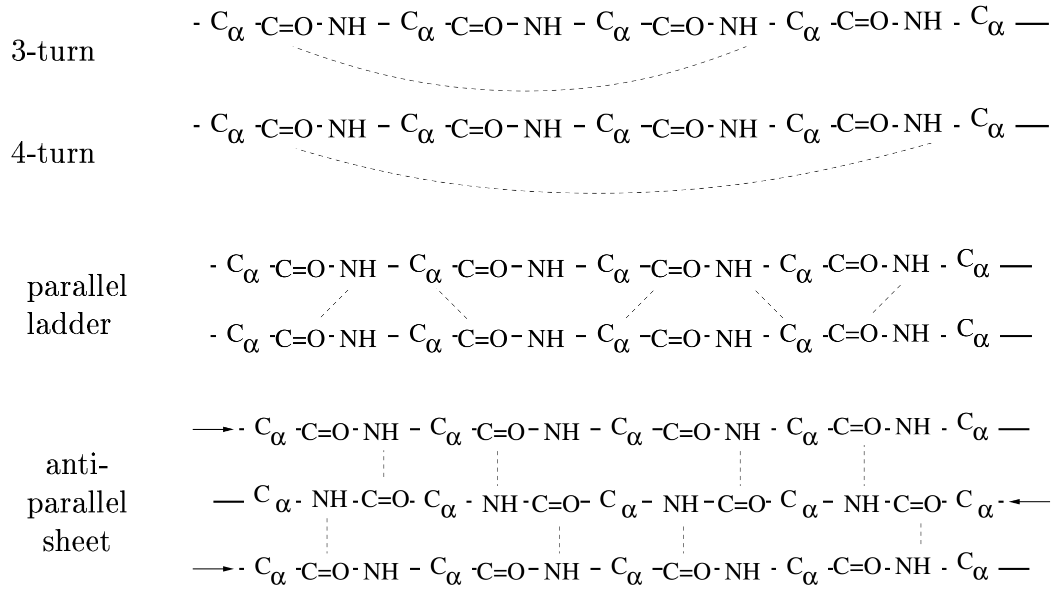
A β-sheet corresponds to a sequence of bridges between consecutive residues in two different regions of the chain. More precisely, we need to introduce two types of patterns:

- a *ladder* is a set of one or more consecutive bridges of the same type, and
- a *sheet* is one or more ladders connected by shared residues.

To allow for irregularities, *β-bulges* are introduced, in which two perfect ladders or bridges can be connected through a gap of one residue in on one side and four on the other.

7.14 Types of H bond patterns

In summary, here are some of the patterns that are used to identify secondary structure elements:


3-turn
- Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα —
4-turn
- Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα —
parallel ladder
- Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα —
- Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα —
anti-parallel sheet
- Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα —
— Cα - NH-C=O- Cα - NH-C=O- Cα - NH-C=O- Cα - NH-C=O- Cα -
- Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα -C=O-NH - Cα —

7.15 DSSP vs STRIDE

The STRIDE algorithm (Frishman and Argos, 1995) extends the DSSP algorithm by adding information on backbone torsion angles. The H-bonding criterion uses a refined energy function and experimental data on H-bond geometry is also taken into account.

In an experimental study of 226 proteins, STRIDE assigned 58% of the proteins closer to the experimental assignment compared to DSSP, whereas DSSP assigns only 31% percent closer than STRIDE. For 11% of the proteins, both computational assignments are in the same proximity of the experimental ones.

7.16 Secondary structure prediction from sequence

Above we discussed how to determine the secondary structure from the coordinates of a known tertiary structure. Now we discuss the more difficult problem of predicting the secondary structure of a protein from its sequence of amino acids. This problem has been studied for more than 30 years and still is an area of active research.

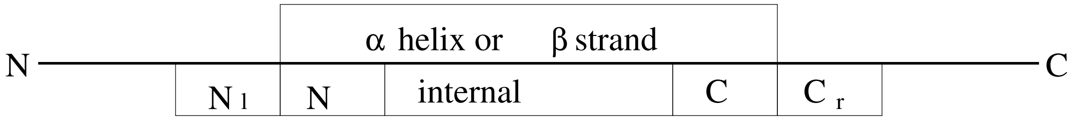
Many different types of approaches have been considered, namely ones: (1) based on stereochemistry rules, (2) using a combination of such rules and statistics, (3) that apply physical models of secondary structure formation, (4) based on statistical information and information theory, (5) that analyze sequence patterns, *(6) based on discriminant analysis*, *(7) using a* *neural net approach*, (8) analyzing evolutionary conservation in multiple alignments, that apply a nearest neighbor analysis, and (10) based on consensus prediction.

7.17 SSP and discriminant-analysis

The definition of secondary structure is a somewhat imprecise and different approaches will determine slightly different structures. For the purpose of modeling tertiary structures, it is less important to assign each individual residue to the correct type of secondary structure than it is to get the location of entire α helices and β strands correct.

The *secondary structure prediction program (SSP)* developed by Solovyev and Salamov (1991, 1994) is aimed at solving this problem.

The SSP algorithm is based on the assumption that secondary structures can be identified by statistical properties of five regions associated with an α-helix or β-strand, namely the N$_l$ region, N-terminal, internal, C-terminal and C$_r$ regions, respectively, as indicated here:



This subdivision into five regions was suggested by experimental data. Additionally, it makes sense to distinguish between short and long structures. I.e., the goal is to predict short and long α helices and β strands.

The SSP algorithm uses a linear combination of three discriminant functions (LDF) to determine whether any given segment of a protein sequence has a preference for being a short or long α helix, or not, and whether it has a preference for being a short or long β strand, or not. If two α helix and β strand predictions overlap, then the one with the larger LDF scores is chosen.

7.18 The singleton characteristic

The *singleton characteristic* is an average of single-residue preferences. Using a database of known protein structures, for every amino acid a the *preference* of being in a specific segment of type k (e.g., the N$_l$ segment of a short α helix or the internal segment of a long β strand) is calculated as

$$S^k(a) = \frac{P^k(a)}{P(a)},$$

where $P(a)$ and $P^k(a)$ are the proportions of amino acids of type a that are contained in the whole database and in segments of type k, respectively (see P.Y. Chou and G.D. Fasman 1978).

Given an amino acid sequence $x = (x_1, \ldots, x_L)$.

Choose start and end positions p and q in the sequence, and a structure type k (e.g, short α-helix). The singleton characteristic $\mathcal{S}^k(p,q)$ is defined as:

$$\mathcal{S}^k(p,q) = \frac{1}{L}\left(\sum_{i=p-m}^{p-1} S_i^{N_l} + \sum_{i=p}^{p+m-1} S_i^{N} + \sum_{i=p+m}^{q-m} S_i^{internal}\right.$$

$$\left. + \sum_{i=q-m+1}^{q} S_i^{C} + \sum_{i=q+1}^{q+m} S_i^{C_r}\right),$$

where $S_i^k := S^k(x_i)$ denotes the preference of amino acid x_i to be contained in a segment of type k.

Here, m is a pre-chosen parameter that determines the size of the non-internal segments N_l, N, C and C_r. It is usually equals 4 or 3 for long α helices and β strands, and $\lfloor\frac{q-p+1}{2}\rfloor$ for short ones.

7.19 Using the singleton characteristic

How do we obtain predictions using the function $\mathcal{S}^k(p,q)$?

Training Present databases contain over 500 independent secondary structures. These can be used as a training set.

We compute the value of $\mathcal{S}^k(p,q)$ for each annotated secondary structure k associated with a subsequence $(x_p,\dots,x_q)$. This gives us a range of values corresponding to correct predictions.

Similarly, for each type of secondary structure k we compute $\mathcal{S}^k(p,q)$ for pairs of positions that do not correspond to annotated structures, obtaining a range of values for incorrect predictions.

Discrimination For each type of secondary structure k a threshold c^k is determined that is used to discriminate between true predictions and false ones. For each pair of positions p and q a prediction of type k for the corresponding subsequence $(x_p, x_{p+1},\dots,x_q)$ is reported, if and only if $\mathcal{S}^k(p,q) > c^k$.

7.20 The doublet characteristic

The *doublet characteristic* is similar to the singlet characteristic. The hope is to obtain a better discrimination by considering *pairs* of amino acids separated by $d = 0, 1, 2$ or 3 other residues.

The *preference* for a particular type of secondary segment k for a pair of amino acids of type a and b, separated by d other residues, is defined as:

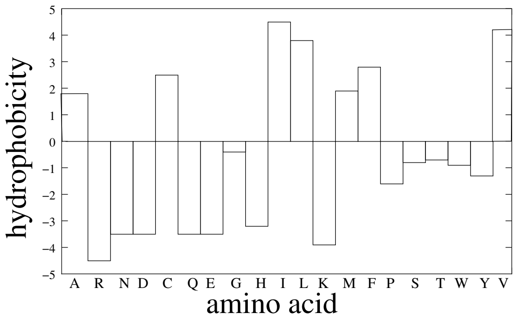
$$D^k(a,b,d) = \frac{P^k(a,b,d)}{P(a,b,d)},$$

where $P(a,b,d)$ and $P^k(a,b,d)$ are the proportions of amino acids of type a and b separated by d others, that are contained in the whole database and in segments of type k, respectively.

The average preference of a segment $(x_p,\ldots,x_q)$ to be in a particular secondary structure k is denoted by $\mathcal{D}^k(p,q,d)$ and is obtained as the sum of all the pair characteristics occurring in the N_l, N, internal, C and C_r segments.

7.21 The hydrophobic moment

Secondary structure prediction can be aided by examining the periodicity of amino acids with hydrophobic side chains in the protein chain. Tables assigning a *hydrophobicity* value $h(a)$ to each amino acid a (Kyte and Doolittle 1982) are used to the determine the hydrophobicity of different regions of a protein:



For example, there is a tendency of hydrophobic residues located in α helices on the surface of a protein to face the core of the protein and for polar and charged amino acids to face the aqueous environment on the outside of the α helix.

This is captured using the concept of *hydrophobic moment*, which is the hydrophobicity of a peptide measured for different angles of rotation per residue (from $0 - 180^o$), giving a measure of the probability that the peptide separates hydrophobic and hydrophilic regions (Eisenberg *et al*, 1984).

For any given segment $(x_p,\ldots,x_q)$ of sequence, the hydrophobic moment for the angle ω is defined as:

$$\mathcal{M}^\omega(p,q) = \left[\left(\sum_{i=p}^{q} h(x_i)\cos(i\omega)\right)^2 + \left(\sum_{i=p}^{q} h(x_i)\sin(i\omega)\right)^2\right]^{\frac{1}{2}},$$

where $h(a)$ denotes the hydrophobicity of the amino acid a.

In the context of predicting α helices and β sheets, the angles of interest are $\omega = 100°$ and $\omega = 160°$, respectively. We use $\omega(k)$ to denote the angle associated with the structure type k.

7.22 Combining the discriminant functions

The SSP method for secondary structure prediction uses a linear combination of all three described discriminant functions (*LDF*, linear discriminating function):

$$\mathcal{Z}^k(p,q) = \alpha_1^k \times \mathcal{S}^k(p,q) + \alpha_2^k \times \mathcal{D}^k(p,q,d) + \alpha_3^k \times \mathcal{M}^{\omega(k)}(p,q).$$

Given a threshold c^k, this function classifies a segment of sequence (x_p, x_q) into *class 1* (i.e., is structure of type k), if $\mathcal{Z}^k(p,q) > c^k$, or *class 2* (i.e., is not structure of type k), if $\mathcal{Z}^k(p,q) \leq c^k$. (Moreover, d is given).

For each type of structure k, the method of *linear discriminant analysis* is used to to determine the coefficients $(\alpha_1^k, \alpha_2^k, \alpha_3^k)$ and the threshold constant c^k. Given a training set, the goal is to maximize the ratio of the between-class variation of $\mathcal{Z}^k$ to within-class variation. (We will skip the details.)

7.23 The SSP algorithm

Given a protein sequence $x = (x_1, \ldots, x_L)$, the SSP algorithm predicts secondary structures in the following way:

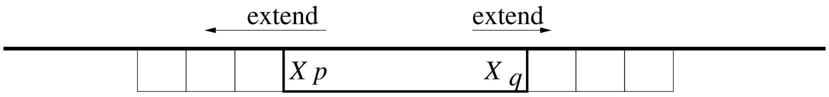
First, a nucleus of a potential α helix is searched for as a segment $(x_p, \ldots, x_q)$ of five residues with an average singleton characteristic higher than a pre-given threshold.

Then, the value of the LDF $\mathcal{Z}^k(p,q)$ for $k =$"short α helix" is compared with the corresponding threshold c^k. If the threshold is not exceeded, then the given segment is deemed unpromising and a new segment is considered.

In the case that the threshold is exceeded, the segment is repeatedly extended in either direction by one residue per step, up to a maximal extension of 15 residues in either direction.

After each extension, the value of the appropriate LDF is computed. The extended segment that gives rise to the highest LDF score is then considered a potential α helix.

A similar *seed-and-extend* strategy is used to determine potential β strand segments:



The result is a collection of potential α helices and β strands.

To obtain a final prediction, overlapping pieces are assigned to the secondary structure types that have the highest LDF value in the region of the overlap. Non-overlapping remainders of such pieces with lower LDF values are retained as predictions, if they are still long enough.

The minimum length for assigning an α helix or β strand is five or three residues, respectively.

SSP server (possibly dead?):
`http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html`

7.24 Measuring prediction accuracy

The accuracy of computational methods that need to be trained on a database of solved structures is often assessed using the *jackknife* procedure: The method is trained on *all but one* data set and then applied to the left out data set D, comparing the result to the true structure associated with D.

To evaluate the performance of a secondary structure prediction, one possibility is to assess the level of single-residue accuracy. For example, this can be measured as the percentage of correct residue predictions Q, or the sensitivity and specificity of the method.

However, this may be problematic, for example, a clearly wrong prediction such as $\alpha\beta\alpha\beta\alpha\ldots$ in an α helix region will still give rise to a score of 50% correct residue predictions.

Thus, in practice one also evaluates the number of correctly predicted α helices and β strands, considering a structure to be *correctly predicted*, if it contains more than a pre-defined number of correctly predicted residues, often just 2.

7.25 Performance of different characteristics

An experimental evaluation of secondary structure predictions was performed on 126 non-homologous proteins with known three-dimensional structures (Rost and Sander 1993), the secondary structure of which was assigned using the DSSP program.

Different combinations of characteristics were compared with each other, giving rise to the following results:

Characteristics used	Q (%)
Singleton characteristic ($\mathcal{S}$)	58.5
$\mathcal{S}$ + hydrophobic moment $\mathcal{M}$	61.4
Doublet characteristic ($\mathcal{D}$)	62.2
$\mathcal{D} + \mathcal{M}$	64.8
$\mathcal{S} + \mathcal{D} + \mathcal{M}$	65.1

(Source: T. Jiang, Y. Xu and M.Q. Zhang (editors), 2002, page 383)

7.26 Segment prediction accuracy

As stated above, single-residue accuracy measures sometimes give a misleading impression of the usefulness of a prediction.

For example, assigning the coil state to all positions in the protein *4sgb* produces a score of $Q = 76.7\%$, although this protein contains several (missed) β structures.

On the other hand, for the protein *3b5c*, SSP correctly predicts four out of five existing α helices and three out of five existing β strands, although the attained Q value is only 56.5%.

The segment prediction accuracy can be estimated as follows: a structure is considered correctly predicted if it has at least two amino acids in common with the correct one.

Under this measure, it was observed that long structures are better predicted than short ones: 89% of α helices of length ≥ 8 and 71% of β strands of length ≥ 6 were correctly predicted, with a specificity of 0.82 and 0.78, respectively.

7.27 A more elaborate discriminant approach

The slightly more recent DSC method (King and Sternberg 1996) is also based on linear discriminant functions and involve from 10 to 27 protein features. Additional post-processing steps are performed to account for higher order properties that cannot be captured by a linear discriminant function. The performance of this method was measured to be 70.1% correctly classified residues.

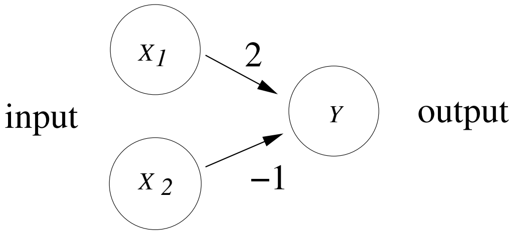
7.28 Neural networks

Several algorithms using neural networks for secondary structures have been developed. One of the most popular is the PHD algorithm by Rost and Sander (1993), which we discuss in detail below.

A *neural network* is a graph whose nodes represent simple processing units and whose (directed) edges define interconnections between the processing elements. Each connection has a *strength*, or *weight*, and the processing ability of the network resides in the weights of its connections. These weights are usually set in a process of adaptation to, or learning from, a given training set.

Neural networks are inspired by neurons in the brain. Neural networks are used for classification problems for which there exist a good supply of training data and little understanding of the structure of the problem at hand.

Here is a very simple example of a neural net whose task it is to determine whether $x_1 > \frac{1}{2}x_2$:



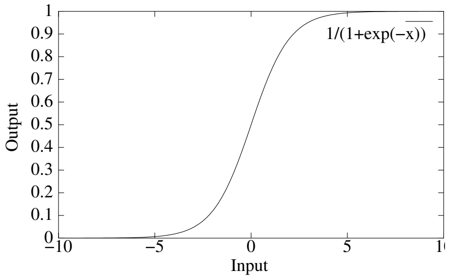
It takes two numbers x_1 and x_2 as input and produces a signal $y = 2x_1 + (-1)x_2$ as output that is positive, if $2x_1 > x_2$, and negative, if $2x_1 < x_2$.

To mimic the *firing* of a neuron, we would like the output of the node labeled y to be 1, if $2x_1 > x_2$ and 0, if $2x_1 < x_2$. This could be realized using a simple step function

$$y = \begin{cases} 1 & \text{if } 2x_1 > x_2 \\ 0 & \text{else.} \end{cases}$$

However, it is better to use a continuous function for this purpose, such as a step-like sigmoidal function of the form:

$$y = \frac{1}{1+\exp(-x)}.$$



In a neural net, a node y that is fed from r nodes $x_1, \ldots, x_r$ by edges (x_i, y) with weights w_i "processes" these inputs and fires a signal of strength $\frac{1}{1+\exp(-x)}$, where $x = \sum_{i=1}^{r} w_i x_i$.

(This is all in the simple case of a feed-forward network, which is what we will consider here.)

7.29 Constructing a neural network

There are two steps to constructing a neural network.

The first step is to design the topology of the network. This involves determining the number of *input nodes* and *output nodes* and how they are associated with external variables. Additionally, the number of internal or *hidden* (layers of) nodes must be determined. Finally, nodes have to be connected using edges.

The second step is called *training. Supervised training* requires a training set consisting of input data points for which the desired output is known. Each such data point is presented to the neural net and then the weights in the net are slightly modified using a gradient descent method so as to increase the performance of the network (as discussed below).

The goal is to set the weights of the edges so that the number of correct results produced for a given training data set is maximized.

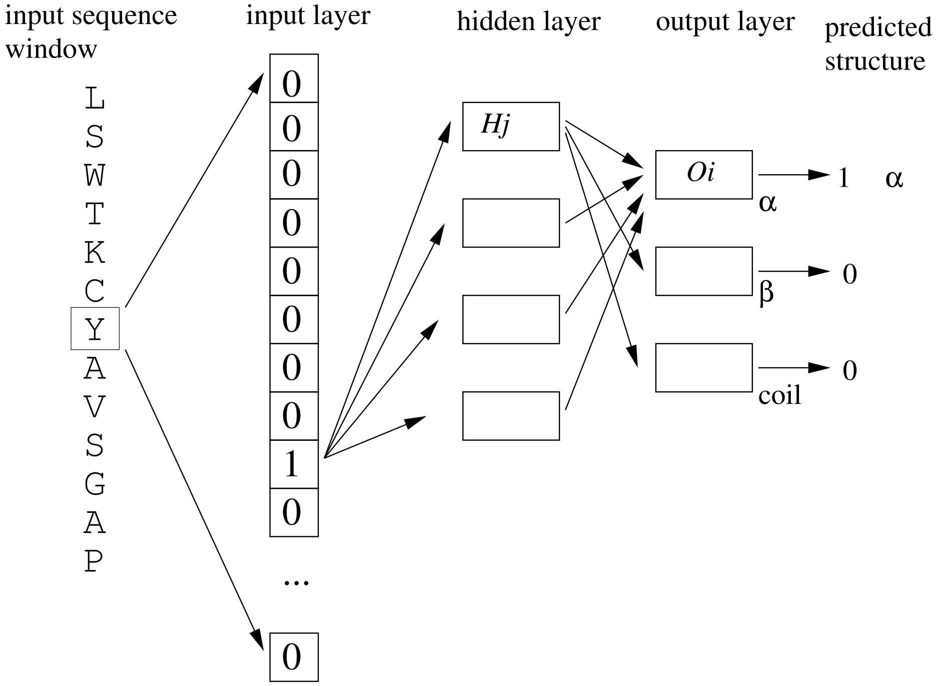
7.30 The PHD neural network

The PHD algorithm by Rost and Sander (1993) uses a neural network to predict the secondary structure of a given residue.

The model consists of three processing units, the input layer, the output layer and a hidden layer. The units of the input layer the amino acids read a small segment (13-17 residues) of sequence around the position of interest, obtained using a *sliding window.*

There are 21 input units per sequence position, namely one per amino acid and one for padding at the beginning and end of the sequence. Given a single sequence, the input unit corresponding to a given amino acid at a given position is set to 1.

Then signals are sent to units in the hidden layer, which process them and pass them on to the units of the output layer. The final output determines which of the three types of secondary structure is assigned to the central residue.



(Adapted from Mount, 2001)

If the input to the neural net consists of a sequence profile, then each input unit is set to the frequency of the associated amino acid at the given position. Additionally, two input units are used to count insertions and deletions.

The predictions obtained for adjacent windows are then post-processed by applying rules or additional neural nets to obtain a final prediction.

Experimental studies show that the PHD method applied to sequences obtains a single-residue accuracy of 70.8%. Application to sequence profiles gives rise to an accuracy of 72% (Rost and Sand 1994).

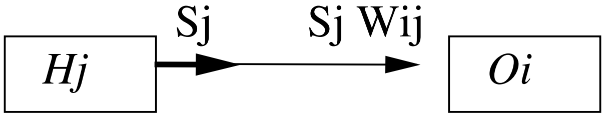
The PHD algorithm uses sequences from the HSSP (homology-derived secondary structure of proteins) database for training (Sander and Schneider, 1991).

7.31 Training the PHD neural network

A method called *back-propagation* can be used to train such neural networks. For example, consider the output node O_i shown in the network above and assume that it predicts whether the central residue lies in an α helix. The output signal s_i predicts an α helix, if it is close to 1, or not, if it is close to 0.

Presented with a training data point, we know whether or not the central residue actually lies in an α helix, and thus, what the desired output d_i of O_i should be.

Consider one of the hidden units H_j that is connected to O_i and emits a signal s_j that is modified by the weight w_{ij}. The signal arriving at O_i is $s_j \times w_{ij}$:



When training the network, the main question is how do we alter w_{ij} so as to bring the value s_i of O_i closer to the desired value d_i? The *gradient descent* method specifies that we modify w_{ij} by the following amount:

$$\Delta w_{ij} = -n\delta E/\delta w_{ij} + m,$$

where the partial derivative of the error E with respect to w_{ij} is given by

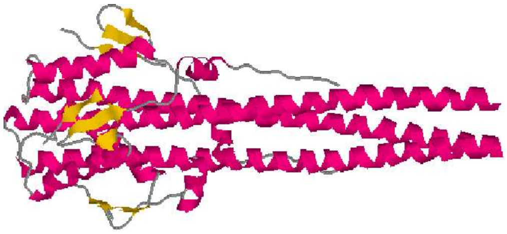
$$\delta E/\delta w_{ij} = (s_i - d_i)s_i(1 - s_i)s_j,$$

and where n is the *training rate* (≈ 0.03) and m is a *smoothing factor* that allows a carryover of a fraction of previous values of w_{ij} (≈ 0.2).

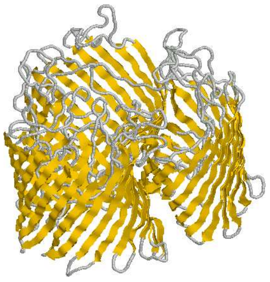
PHD web server:
`http://www.embl-heidelberg.de/predictprotein/predictprotein.html`

7.32 Configurations of secondary structure elements

Secondary structure elements such as α helices and β sheets can sometimes group into larger structures such as (a) a *coiled coil*, a configuration in which α helices (originally called "coils") are wound into a superhelix, or (b) a *parallel β helix* in which β strands wrap around to form a helix.



(a) coiled coil
Hemagglutinin at pH 4.0
1HTM



(b) parallel β helices
Maltoporin Sucrose Complex
1AF6

7.33 Coiled coils

In the following we discuss how to predict coiled coils from protein sequence. This is mainly based on the following two papers:

- Andrei Lupas, Marc van Dyke and Jeff Stock, *Predicting coiled coils from protein sequences*, Science, 252:1162-64 (1991), and
- Andrei Lupas, *Coiled coils: new structures and new functions*, TIBS 21:375-382 (1996).

Coiled coils were first described in 1953 by Pauling and Corey, and, independently, by Crick, as the main structural element of a large class of fibrous proteins that included keratin, myosin and fibrinogen.

About three percent of all proteins are thought to contain a coiled coil domain. Hence, this type of configuration is probably important for many cellular processes.

7.34 Description of coiled coils

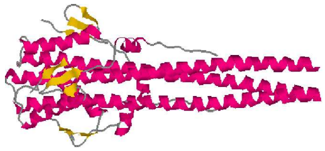
By definition, *coiled coils*

- are formed by two or three α helices in parallel in and register that cross at an angle of $\approx 20°$,
- are strongly amphipathic and display a pattern of hydrophilic and hydrophobic residues that is repeated every seven residues, and
- their sequences exhibit common patterns of amino acid distribution that appear to be distinct from those of other proteins.

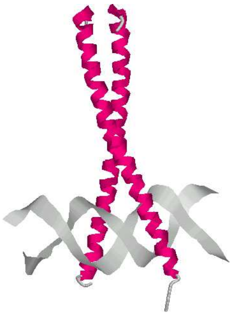
The prediction of coiled coil domains from protein sequence is based on the two latter observations. Based on them, it can be predicted with significant reliability which α helices participate in a coiled coil. However, if more than two such α helices are present, it is usually very difficult to predict *which ones will match up* to form a specific coiled coil.

Here are some examples of coiled coils:

Tropomyosin (2TMA)



Hemagglutinin
at pH 4.0 (1HTM)



GCN4 with
DNA-binding domain (1YSA)

The *leucine zipper* domain is typically made of two anti parallel α helices held together by interactions between hydrophobic leucine residues located at every seventh position in each helix, see 1YSA above. The zipper holds protein subunits together.

In the transcription factors Gcn4, Fos, Myc, and Jun, the binding of the subunits form a scissor-like structure with ends that lie on the major groove of DNA.

Coiled coils fulfill a variety of functions: they can form large, mechanically rigid structures, e.g. hair or feathers (keratin), or blood clots (fibrin), the cellular skeleton (intermediate filaments), provide a scaffold for regulatory complexes (tropmyosin), form spacers that separate the outer membrane from the cell wall in bacteria (murein lipoprotein), and provide a protective surface for pathogens (the M proteins of staphylococci). (Lupas 1996)

7.35 The heptad repeat

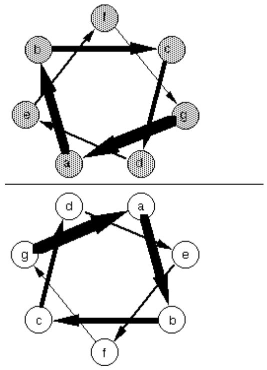
In an α helix, it takes about 3.62 amino acids to complete one turn of the helix and so the positions of the residues around the central axis of an α helix do not display a short periodicity.

However, if a right-handed α helix is given a slight left-handed twist, then the number of residues per turn can be reduced to 3.5 and the positions will display a periodicity of 7.

(By twisting the helix in the other direction to about 3.7 residues per turn, a periodicity of 11 can be achieved, but it is unclear whether such right-handed coiled coils actually exist.)

In a coiled coil configuration, the participating α helices do indeed give each other a slight left-handed twist, thus enabling themselves to line up along a periodic subset of amino acids.

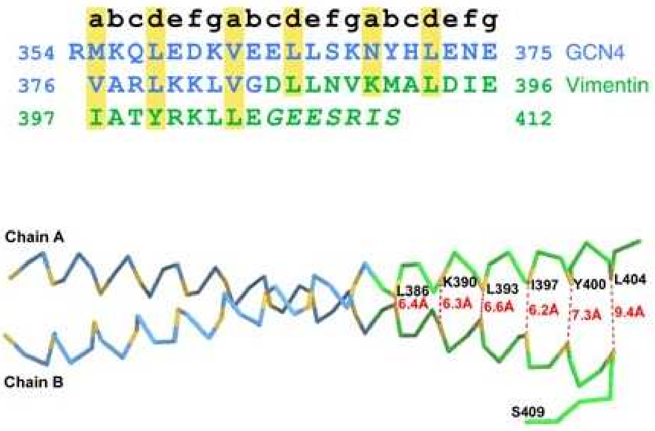
Viewed from above, the configuration of two α helices forming a coiled coil can be displayed using a *helical-wheel* plot:



http://www.cryst.bbk.ac.uk/PPS95/course/6_super_sec/cc.html

The seven periodic classes of amino acid positions are called a, b, c, d, e, f and g. The positions a and d are filled by hydrophobic residues and form the helix interface. The other residues are hydrophilic and form the solvent exposed part of the coiled core.

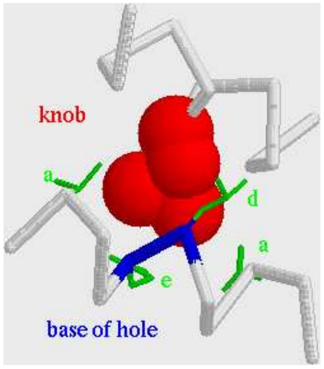
Here is another illustration of the periodic nature of the distribution of hydrophobic residues along α helices participating in coiled coils:



http://www.biozentrum.unibas.ch/personal/burkhard/

7.36 Knobs into holes

In this configuration, the residues of the two helices mesh nicely in what is called a *knobs-into-holes* packing:



http://www.cryst.bbk.ac.uk/PPS95/course/6_super_sec/cc.html

Here, a residue from one helix (knob) packs into a space surrounded by four side chains of the facing helix (hole).

7.37 Obtaining coiled-coil statistics

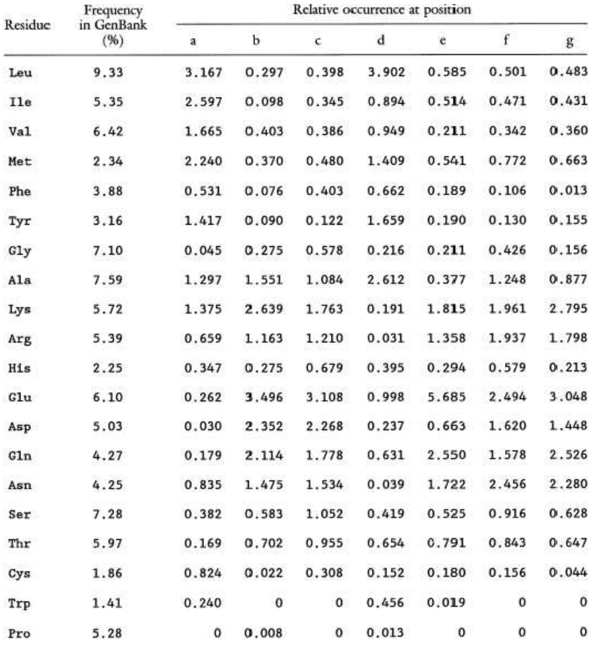
To be able to predict coiled coil forming α helices from sequence, a database of training sets is needed.

The sequences of coiled-coil domains from tropmyosins, myosins, and keratins deposited in GenBank provide a coiled-coil database.

For each of the twenty amino acids A, one can determine the frequency $P_i(A)$ with which it occurs at position $i \in \{a, \ldots, g\}$ of the heptad repeat and the frequency $P(A)$ with which A occurs anywhere, in any sequence in GenBank. This then gives rise to the *relative frequency* with which A occurs at position i:

$$F_i(A) = \frac{P_i(A)}{P(A)}.$$

Relative frequencies reported by Lupas et al (1991):



Residue	Frequency in GenBank (%)	Relative occurrence at position						
		a	b	c	d	e	f	g
Leu	9.33	3.167	0.297	0.398	3.902	0.585	0.501	0.483
Ile	5.35	2.597	0.098	0.345	0.894	0.514	0.471	0.431
Val	6.42	1.665	0.403	0.386	0.949	0.211	0.342	0.360
Met	2.34	2.240	0.370	0.480	1.409	0.541	0.772	0.663
Phe	3.88	0.531	0.076	0.403	0.662	0.189	0.106	0.013
Tyr	3.16	1.417	0.090	0.122	1.659	0.190	0.130	0.155
Gly	7.10	0.045	0.275	0.578	0.216	0.211	0.426	0.156
Ala	7.59	1.297	1.551	1.084	2.612	0.377	1.248	0.877
Lys	5.72	1.375	2.639	1.763	0.191	1.815	1.961	2.795
Arg	5.39	0.659	1.163	1.210	0.031	1.358	1.937	1.798
His	2.25	0.347	0.275	0.679	0.395	0.294	0.579	0.213
Glu	6.10	0.262	3.496	3.108	0.998	5.685	2.494	3.048
Asp	5.03	0.030	2.352	2.268	0.237	0.663	1.620	1.448
Gln	4.27	0.179	2.114	1.778	0.631	2.550	1.578	2.526
Asn	4.25	0.835	1.475	1.534	0.039	1.722	2.456	2.280
Ser	7.28	0.382	0.583	1.052	0.419	0.525	0.916	0.628
Thr	5.97	0.169	0.702	0.955	0.654	0.791	0.843	0.647
Cys	1.86	0.824	0.022	0.308	0.152	0.180	0.156	0.044
Trp	1.41	0.240	0	0	0.456	0.019	0	0
Pro	5.28	0	0.008	0	0.013	0	0	0

7.38 Sliding window evaluation

Given a protein sequence $x = (x_1, \ldots, x_L)$. These relative frequencies $F_i(A)$ are used for prediction as follows:

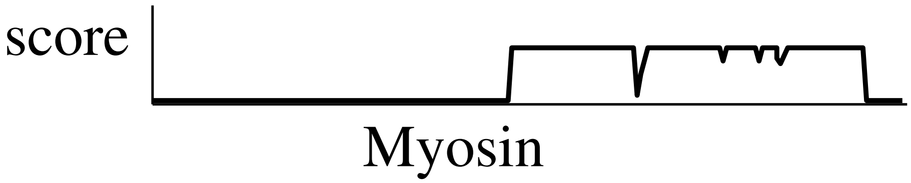
A *sliding window* of length 28 residues is moved along the sequence and for each start position $p = 1, 2, \ldots, L-27$ of the window, the following steps are performed:

- the window is assigned a heptad repeat frame,
- each residue in the window is assigned the appropriate frequency f_i obtained from the above table, and then
- the geometric mean G all these values $f_1, f_2, \ldots, f_{28}$ is computed, $G = \left(\prod_{i=1}^{28} f_i\right)^{\frac{1}{28}}$.

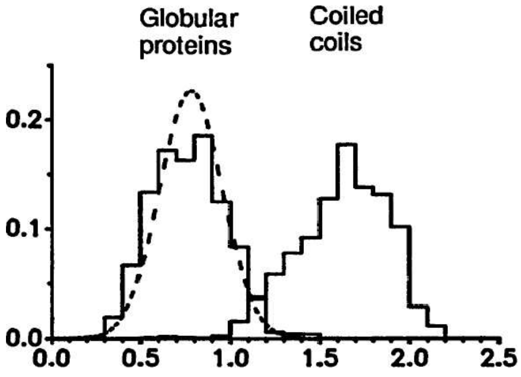
Thus, each choice of window and heptad repeat frame is given a score.

Consider a fixed residue x_i: it is contained in 28 different windows and for each such window, there are 7 different choices for heptad frames. (If the residue is close to one end of the sequence, then this number is smaller, of course.) The residue x_i is assigned a score that is simply the largest score assigned to any window (and heptad repeat frame) that contains it.

Because the maximal score over all windows is taken, we obtain a step-like score function along the sequence, e.g.:



Running the algorithm on a collection of globular proteins and then on a collection (of roughly the same size) of known coiled coils produces the following distribution of scores:



Lupas et al. 1991

For globular proteins, the mean score is 0.77 with a standard deviation of 0.20. For coiled-coil sequences, the mean score is 1.63 and the standard deviation is 0.24.

7.39 Estimation of probability of being a coiled coil

The above score distributions allows an estimate of the probability that a residue with a given score would be in a coiled score. The ratio of globular to coiled-coil proteins is estimated to be approximately 1 : 30. The probability P of forming a coiled coil of a given score S is then:

$$P(S) = \frac{G_{cc}(S)}{30G_g(S) + G_{cc}(S)},$$

where G_g and G_{cc} are two Gaussian curves that approximate the distribution of globular and coiled-coil sequences, respectively. This probability is then used to predict coiled coils.

7.40 Implementation

An implementation of the approach described here can be run at: `http://www.ch.embnet.org/software/COILS_form.html`.