

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:

$$\Delta w_{ij} = -n \partial E / \partial w_{ij},$$

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

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

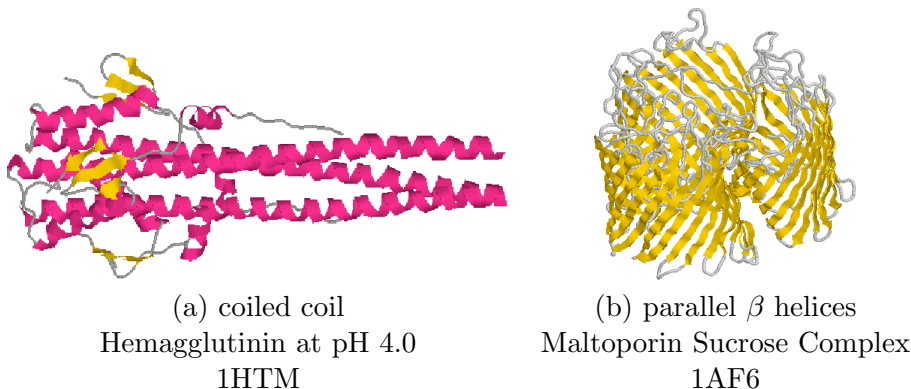
and where n is the *training rate* (≈ 0.03).

Updated PHD web server:

<http://www.predictprotein.org/>

9.8 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.



9.8.1 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.

9.8.2 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 $\approx 20^\circ$,
- are strongly amphipathic and display a pattern of hydrophilic and hydrophobic residues that is repeated every seven residues, and

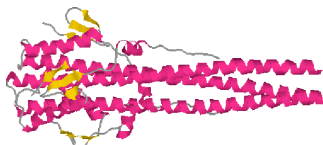
- their sequences exhibit common patterns of amino acid distribution that appear to be distinct from those of other proteins.

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.

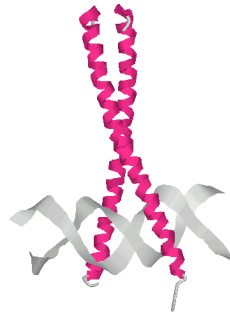
Here are some examples of coiled coils:



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 (tropomyosin), 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)

9.8.3 The heptad repeat

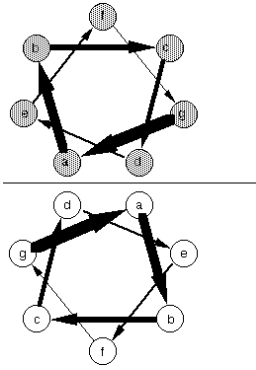
In an α helix, it takes about 3.62 amino acids to complete one turn of the helix and so the positions of the residues around the central axis of an α helix do not display a short periodicity.

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

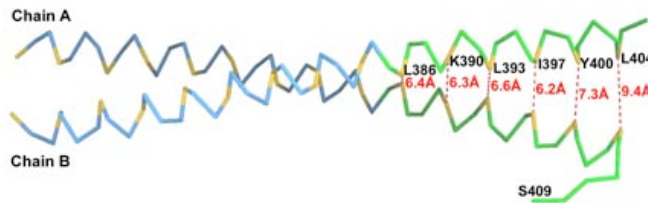
The seven periodic classes of amino acid positions are called *a*, *b*, *c*, *d*, *e*, *f* and *g*. The positions *a* and *d* are filled by hydrophobic residues and form the helix interface. The other residues are hydrophilic and form the solvent exposed part of the coiled core.

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

```

          abcdefgabcdefgabcdefg
354  RMKQLEDKVEE LLSKNYHLENE 375  GCN4
376  VARLKKLVGD LLNVKMALDIE 396  Vimentin
397  IATYRKLL EGEESRIS      412

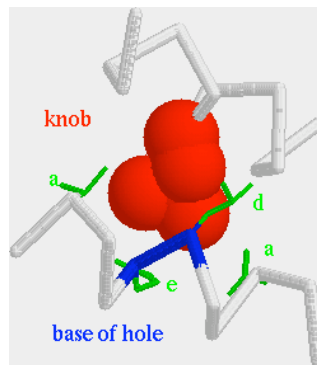
```



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

9.8.4 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).

9.8.5 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 tropomyosins, 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 $i \in \{a, \dots, g\}$ 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

9.8.6 Sliding window evaluation

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

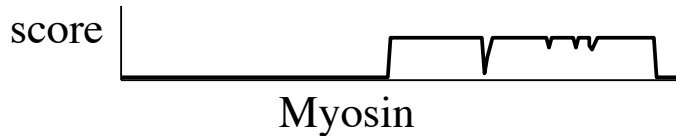
A *sliding window* of length 28 residues is moved along the sequence and for each start position $p = 1, 2, \dots, L - 27$ of the window, the following steps are performed:

- 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, \dots, f_{28} is computed, $G = \left(\prod_{i=1}^{28} f_i \right)^{\frac{1}{28}}$.

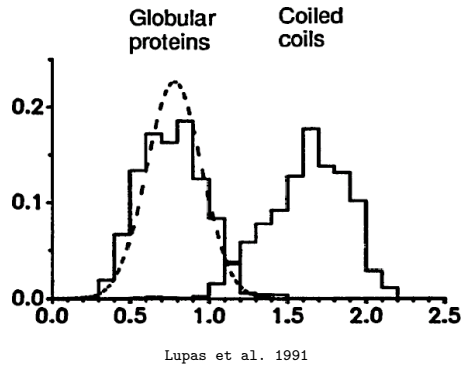
Thus, each choice of window and heptad repeat frame is given a score.

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:



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.

9.8.7 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.

9.8.8 Implementation

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