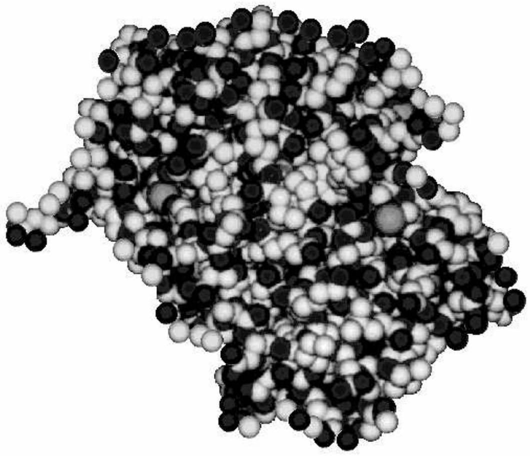
Predicting protein secondary structure

A protein is basically a string of amino acids (or **residues**). Predicting the three-dimensional shape of proteins from their amino acid sequence is widely believed to be one of the hardest unsolved problems in molecular biology. It is also of considerable interest to pharmaceutical companies since shape generally determines the function of a protein. This is what a protein looks like:



The sequence of amino acids is called the **primary structure** of the protein. Spatially, the amino acids are arranged in different patterns (spirals, turns, flat sections etc.). The interest here is in one particular shape (or **secondary structure**), namely when the protein shape is a helical one. This shape is called an α-helix.

The data here is from the *Golem* [1] experiments reported in [2]. The task is to learn rules to identify whether a position in a protein is in an α-helix. Points of relevance are:

- Each amino acid is denoted by a lower case character. There are 20 such amino acids.
- Positive examples state which positions of chosen proteins are in an α-helix. Negative examples state the positions that are not in an α-helix.
- The following background knowledge is provided:
  - **position(A,B,C)**. Residue of protein A at position B is C.
  - **octf(A,B,C,D,E,F,G,H,I)**. Arithmetic information the allows indexing groups of nine adjacent positions in a protein. Basically says positions A–I occur in sequence.

- **alpha_triplet(A,B,C)**. Arithmetic information the allows indexing groups of three adjacent positions in a protein.
- **alpha_pair(A,B)**. Arithmetic information the allows indexing a pair of adjacent positions in a protein.
- **alpha_pair4(A,B)**. Arithmetic information the allows indexing a pair of positions separated by 4 positions in a protein.
- Physical and chemical properties of individual residues are described by unary predicates. These properties include hydrophobicity, hydrophilicity, charge, size, polarity, whether a residue is aliphatic or aromatic, whether it is a hydrogen donator or acceptor etc.
- Sizes, hydrophobicities, polarities etc., are represented by constants (for example, polar0, polar1 etc.).
- Relations between the constants (less_than(polar0,polar1)) is also provided as background knowledge.

The data files are contained within one compressed TAR file. Within that, they are as used in the original *Golem* experiments. That is, background knowledge files have a ".b" suffix, positive example files have a ".f" suffix, and negative example files have a ".n" suffix.

References

[1] Muggleton S.H. and Feng C. (1990). Efficient induction of logic programs. *Proceedings of the First Conference on Algorithmic Learning Theory*, Tokyo.

[2] Muggleton S., King R.D., and Sternberg M.J.E. (1992). Predicting protein secondary structure using inductive logic programming. *Protein Engineering*, 5:647–657.