1 Introduction

Proteins play essential roles in almost all biological processes. Their wide range of activities arises from the variety of three-dimensional structures they can adopt. Therefore, understanding protein structure is one of the major challenges of molecular biology. Despite more than three decades of research, the goal of predicting the three-dimensional structure of a protein from the knowledge of sequence information alone remains elusive. However the explosion of sequence data is now putting tremendous pressure for progress to be made.

The number of known three-dimensional structures, determined through X-ray crystallography and Nuclear Magnetic Resonance experiments, is also increasing rapidly. There are now approximately 10,000 protein structures in the public repository. To ease understanding classification schemes have recently been developed. One example is SCOP (Structural Classification of Proteins) [3]. The schemes are hierarchical, proteins which are known to have evolved from a common ancestry are grouped together into families, and super-families. The next level puts together proteins that share the same fold, *i.e.* the same core secondary structure elements and the same interconnections. In this case, the similarity may be the result of convergence towards a stable architecture. At this level, the proteins have quite dissimilar sequences which makes it impossible for sequence comparison methods to detect the relationship. In this work the SCOP classification scheme is the starting point for a machine learning experiment which aim to relate structural principles to the concept of folds. We present an accessible summary of our latest results of which the details have already been reported in three papers [10, 9, 11].

2 Protein 3D structure

The three-dimensional structure of proteins is highly complex. In general, three levels of abstraction are distinguished: primary, secondary and tertiary structure. Proteins are long chains of amino acids. There are 20 naturally occurring amino acids, each with different chemical properties. The amino acids are linked by a covalent bond to form chains, typically 100 to 500 amino acids long, and referred to as primary structure or sequence. A particular sequence folds into a specific compact three-dimensional or tertiary structure. The two predominant methods to structure determination are X-ray crystallography and NMR spectroscopy. Those techniques require sophisticated equipments and because of technological limitations, the sequences of amino acids are routinely determined in large quantities whilst the determination of the three-dimensional structure remains difficult. Early on it was predicted that segments of the primary sequence would adopt local regular structures [8], the two main types are the α-helices and the β-strands, while the intervening regions are called loops or coils, collectively those elements are referred to as the secondary structure.

Identifying rules which explain the observed folds remains a challenge and often involves manual intervention of experts [3, 2, 7]. For several folds, these signatures are reported in the literature, generally after extensive study. Few experts are familiar with many of these rules and the knowledge is certainly not formalised, with a common language, in a form suitable for automated testing as new structures are determined. Also, automated methods can identify features that are missed by manual examination.

3 Approach

The objective of this work is to automate the discovery of structural rules. Inductive Logic Programming (ILP) is a logic-based approach to machine learning. ILP is particularly well suited to study problems encountered in molecular biology. First, protein structures are the result of complex interactions between sub-structures (secondary structures) and the ability to learn relations might prove to be a key feature. Second, ILP systems can make use of problem-specific background knowledge taking advantage of the vast amount of knowledge that has been accumulated. Third, ILP uses a common representation for the examples, the background knowledge and the hypotheses, and therefore provides a good integration for the development of applications together with the machine learning experiments. Finally, the hypotheses can be made readable, by straightforward translation to natural languages, and integrated to the cycles of scientific debates. In complex domains, such as the structure determination, it is unlikely that a breakthrough will come from a single machine learning experiment, the ability of ILP to make the rules readable is therefore an important advantage to assist the process of scientific discovery.

3.1 Machine learning algorithm

Inductive Logic Programming is concerned with the induction of hypotheses from examples and background knowledge [6]. In this work, we use Progol which is being developed by the second author [5]. As mentioned above, a restricted subset of first-order logic is used as a common representation for the examples, the background knowledge and also the generated hypotheses. In the case of the protein folds problem, a (positive) example represents the fact that the domain `d1hlb__` belongs to the Globin fold by `fold('Globin-like', d1hlb__)`. The background knowledge contains information such as the relationships between secondary structures and the presence of a proline. The algorithm then constructs a hypothesis which explains this example in terms of the background knowledge, the following rule was generated for the Globin-like fold,

```
fold('Globin-like', X) :-
    adjacent(X, _, B, 1, h, h),
    has_pro(B).
```

which is interpreted as "domain `X` belongs to the Globin fold if its first helix is followed by another one that contains a proline".

More specifically, the background knowledge for those experiments contains information about the secondary structure, calculated with PROMOTIF [4] from experimental three-dimensional structures. For each secondary structure we calculate the average hydrophobicity, the hydrophobic moment and the number of amino acids. The presence of a proline is also noted. For each inter-secondary structure region we calculate the number of amino acids. The background knowledge contains global information as well: the total number of strands and helices and the total number of amino acids. Here we also add information about protein topology, such as the packing of helices, the relative direction of β-strands and the types of β-sheets.

4 Results and discussion

In a previous work, we have compared two different representations of the background knowledge [10, 9]. The first contained only predicates which encode global characteristics of protein folds, specifically, the total number of residues and the total number of secondary structures of both types, helices and strands. For the second, new predicates were added which introduce relationships between secondary structure elements and their properties. The results showed that it is possible to construct good classifiers with a background knowledge which is essentially limited to attribute-values. However, higher accuracy figures were obtained with the relational representation. Furthermore, in the case of the relational dataset some of rules can be related to results published in the relevant scientific literature. One such example is that of the Globin fold.

Rule 1 (Globin fold) *Helix A at position 1 is followed by helix B. B contains a proline residue.*

```
fold('Globin-like', X) :-
    adjacent(X, A, B, 1, h, h),
    has_pro(B).
```

A distinctive feature of this fold is the presence of a conserved proline residue in helix B, which causes a sharp bend in the main chain. This observation has been reported previously by Bashford *et al.* [1] and has been rediscovered here.

One of the main limitations of this application concerns the representation. Secondary structure positions are counted from the N-terminal end of the structure and do not take into account the possibility of insertions. We have developed a new representation that 1) sequentially numbers the secondary structures for the C-terminal as well as N-terminal and 2) includes additional information about the topology of the sheets and the packing the helices [11]. Preliminary runs show that Progol can now learn descriptions such as the following:

```
fold(A,'SH3-like barrel') :-
    number_strands(4=<A=<7),
    sheet(A,B,anti),
    has_n_strands(B,5),
    strand(A,C,B,1),
    strand(A,D,B,-1),
    antiparallel(C,D).
```

which allows for insertion into the sheet since the relation `antiparallel(C,D)` is between the first and the last strand. Cross-validation experiments were conducted for the 20 most populated folds. The overall cross-validated accuracy was found to be 75.1 ± 1.6 % for the more limited background knowledge, and 82.1 ± 1.4 % with additional information. The expected accuracy of a random prediction is 50 %.

In terms of biology, the new rules are more descriptive than the previous ones. Prior rules represented local motifs, often associated with functional regions, the new rules represent more complete descriptions similar to comments often found in SCOP classification itself.

Those experiments show that ILP can be used effectively to learn rules in complex domains such as protein structure. The rules produced in the context of the relational learning experiments, were found to be more informative, as judged by our knowledge of protein structure, than those generated in the context of attribute-value experiments. The rules can be explained in terms of structural and/or functional concepts, such active site and binding location. When information about the topology is added the rules are often more descriptives, similar to those descriptions found in SCOP.

Inductive Logic Programming has allowed us to explore several representations and to effectively learn rules in a complex domain such as structural biology.

Acknowledgements

This work is supported by a BBSRC/EPSRC Bioinformatics grant.

References

[1] D. Bashford, C. Chothia, and A. M. Lesk. Determinants of a protein fold. unique features of the globin amino acid sequences. *J. Mol. Biol.*, 196(1):199–216, 1987.

[2] Carl Branden and John Tooze. *Introduction to Protein Structure*. Garland, 2nd edition, 1999.

[3] S. E. Brenner, C. Chothia, T. J. Hubbard, and A. G. Murzin. Understanding protein structure: using SCOP for fold interpretation. *Methods in Enzymology*, 266:635–43, 1996.

[4] E. G. Hutchinson and J. M. Thornton. PROMOTIF – a program to identify and analyze structural motifs in proteins. *Protein Science*, 5(2):212–20, 1996.

[5] S. Muggleton and J. Firth. CProgol4.4: Theory and use. In Sašo Džeroski and Nada Lavrac, editors, *Inductive Logic Programing and Knowledge Discovery in Databases*. 1999. forthcoming book.

[6] S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *J. Logic Programming*, 19/20:629–679, 1994.

[7] C. A. Orengo, D. T. Jones, and J. M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372(6507):631–4, 1994.

[8] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.*, 37:205–210, 1951.

[9] M. Turcotte, S.H. Muggleton, and M.J.E. Sternberg. Automated discovery of structural signatures of protein fold and function. *J. Mol. Biol.*, 306(3):591–605, 2001.

[10] M. Turcotte, S.H. Muggleton, and M.J.E. Sternberg. The effect of relational background knowledge on learning of protein three-dimensional fold signatures. *Machine Learning*, 43:81 – 95, 2001.

[11] M. Turcotte, S.H. Muggleton, and M.J.E. Sternberg. Generating protein three-dimensional fold signatures using inductive logic programming. *Computers & Chemistry*, 2001. *in press*.