Inductive Logic Programming used to Discover Topological Constraints in Protein Structures

Ross D. King[1], Dominic A. Clark[2],
Jack Shirazi[2] and Michael J.E. Sternberg[1]

1 Biomolecular Modelling Laboratory, 2 Biomedical Informatics Unit
Imperial Cancer Research Fund, 44 Lincoln's Inn Fields,
London, WC2A 3PX, U.K, Tel. +44-71-242-0200, Fax.+44-71-269-3417,
rd_king@icrf.ac.uk, dac@biu.icnet.uk, js@biu.icnet.uk, m_sternberg@icrf.ac.uk

Abstract

This paper describes the application of the Inductive Logic Programming (ILP) program GOLEM to the discovery of constraints in the packing of beta-sheets in alpha/beta proteins. These constraints (rules) have a role in understanding the protein folding problem. Constraints were learnt for four features of beta-sheet packing: the winding direction of two sequential strands, whether two consecutive strands pack parallel or anti-parallel, whether two strands pack adjacently, and whether a beta-strand is at an edge. Investigation of the learnt constraints revealed interesting patterns, some of which were previously known, others that were novel. Novel features include the discovery: that the relationship between pairs of sequential strands is in general one of decreasing size, and that more sequential pairs of strands wind in the direction out than the direction in. We conclude that machine learning has a useful place in molecular biology as a pattern discovery tool.

Introduction

Molecular Biology is currently experiencing an enormous increase in the quantity of available data. This has brought into focus the need for automatic tools to aid scientists data discover patterns in data. In this paper we test the application of Inductive Logic Programming (ILP) to the discovery of patterns/rules in molecular biological data. The particular problem studied was the discovery of constraints in the folding patterns of alpha/beta proteins; a part of the protein folding problem .

The protein folding problem is the problem of predicting protein structure from sequence. A common approach to this problem is to exploit the hierarchical nature of protein structure - a divide and conquer strategy. This approach starts with protein secondary structure prediction (Garnier *et al.* 1978; Muggleton *et al.* 1992; Rost & Sander 1993). Success at secondary structure prediction has been limited to a ceiling of about 70%. It is generally believed that the reason for this is that there are constraints on secondary structure imposed by higher levels of protein structural organisation. Such higher levels of structure include the packing together of beta-strands to form beta-sheets, the clustering together of beta-sheets and alpha-helices to form protein domains, etc. In this paper constraints are learnt relating how beta-strands pack together to form sheets. One of the most important sub-types of proteins are those that have only alpha/beta domains. These consist of repeated alpha-helices and beta-strands. Such domains were chosen as a test bed for our application of machine learning. Several factors make these proteins particularly suitable for testing the application of machine learning.

- The relevant data for the problem exists in a form suitable for machine learning. The data consists of Prolog clauses using the symbolic representation scheme developed for the TOPOL database of protein topology (Rawlings *et al.* 1985).
- Several workers have previously investigated patterns in alpha/beta proteins (Sternberg & Thornton 1977a, 1977b; Branden 1980; Richardson 1981; Taylor & Green 1989). The hand crafted patterns described by these workers provide a useful comparison to the machine learning produced rules.
- Clark *et al* (1991, 1993) have taken the existing hand-crafted patterns, checked their validity, and formalised them; in the process additionally formalising some rules of their own. They have used these rules in an algorithm for predicting the domain packing structure of a protein from its secondary structure. The rules are formalised as constraints in the constraint logic programming language (ElipSys). The rules generated by machine learning can be directly tested against the hand crafted rules on the task of predicting protein domain structure.

The ILP program GOLEM (Muggleton & Feng 1990) was chosen as the inductive method for pattern discovery. The reasons for this were: the data already exists in a data format suited to ILP, we consider ILP to be better suited than propositional methods to machine discovery problems in general, and we have successfully applied GOLEM to other related problems in biological structures. The protein data exists in a logic programming format; this data consists of complicated relations that reflect the complicated structures of proteins. These relations would be difficult to represent using attributes. We consider ILP programs more suitable for machine discovery than attribute based ones because: they can better express background knowledge (there is commonly a great deal of relevant background knowledge in scientific discovery problems), and they can learn a more general set of concepts (it is difficult to represent many scientific concepts using attributes). GOLEM has previously been applied to the prediction of protein secondary structure (Muggleton *et al.* 1992), the design of drugs to fit into the active sites of proteins (King *et al.* 1992), and machine discovery (King *et al.*, 1994).

Methods

Database of Proteins

The proteins chosen were of class alpha/beta. The original data comes from the PIPS deductive database (Shirazi & Clark 1993), which is in turn derived from the IDITIS relational database (marketed by Oxford Molecular and developed by Gardner and Thornton extending the work of (Islam & Sternberg 1989)). The PAPAIN database is a deductive database implemented in the constraint logic-programming language ElipSys. The selection of proteins was carried out using the non-homologous set of protein folds described by (Orengo *et al.* 1993) Brookhaven release 61. All protein domains were chosen that are in the IDITIS database and the Orengo sub-classes (80 Str) alpha/beta: Doubly Wound, alpha/beta: One Doubly Wound, and Alpha/beta: Two Doubly Wound. Domains in the Orengo sub-class alpha/beta Tim Barrel were excluded because of problems in the designation of barrels in the IDITIS database. The data was further filtered by removal of homologous beta-sheets, so that for each protein only non-homologous sheets were included. The gave the following set of proteins: 1CSE, 2FCR, 2TRX, 3ADK, 3CHY, 3PGM, - 4DFR, 4FXN, 5CPA, 5P21, 1GD1, 1PHH, 2TS1, 3GRS, 6LDH, 8ADH, 8CAT, 1RHD, 2LIV, 2YHX, 3GBP (*S. Typhimurium:* Orengo used *E. Coli*), 3PGK, 4PFK. From this set the following proteins were randomly chosen as a test set (7 from 23): 3ADK, 3CHY, 4FXN, 1GD1, 8ADH, 1RHD, 3GBP. The remaining proteins were used as a training set.

Data Representation

The ILP program GOLEM was used in this study. GOLEM takes as input: positive examples, negative examples, and background knowledge described as Prolog ground clauses. The Prolog representation used in GOLEM is a modified version of that used in TOPOL (Rawlings *et al.* 1985). It draws on the experience of the representations used to represent protein sequences in (Muggleton *et al.* 1992; King *et al.*, 1994) and drugs in (King *et al.* 1992). In the experiments a basic set of background facts was used. From these facts, a number of different predicates describing alpha/beta proteins were learnt. The basic idea is that the background knowledge consists of facts that could be assumed to be known after a successful secondary structure prediction (Fig. 2). From this information it is possible to predict the next level of protein structure, how the secondary structure elements pack together. The learnt descriptive predicates describing alpha/beta proteins will aid in this prediction of protein structure. For each descriptive predicate, rules were learnt for it and its negation: e.g. predicate foo(A), and not foo(A). The intention was to produce rules that had high accuracy and statistical significance. It was not intended that rules would necessarily cover all the examples, i.e. discriminate between examples with and without the foo predicate. The reasons for this are: the rules are intended to be input into a constraint-satisfaction algorithm, and the rules are intended to be understandable to protein chemists.

The following descriptive predicates were learnt:
edge(P,S).
In protein P strand S is positioned at the edge of a sheet.
parallel(P,S1,S2).
In protein P consecutive strands S1 and S2 are parallel.
adj(P,S1,S2).
In protein P strands S1 and S2 are adjacent in a sheet.
in(P,S1,S2).
In protein P strand S1 is closer to the edge than strand S2.
printed

The following structural background knowledge was used (P is a protein id.).
seql(P,S,L).
In P strand S has length L.
con_helix(P,S1,S2,H).
In P strands S1 and S2 are joined by H helices.
con_length(P,S1,S2,R).
In P strands S1 and S2 are joined by R residues.
con_to_helix1(P,S1,S2,N).
In P between S1 and S2 the first helix has id. N.
con_to_helix2(P,S1,S2,N).
In P between S1 and S2 the last helix has id. N.
con_to_chain1(P,S1,S2,N).
In P between S1 and S2 the first coil has id. N.
con_to_chain2(P,S1,S2,N).
In P between S1 and S2 the last coil has id. N.

The following hydrophobic background knowledge was used (In P,E, means in protein P, sheet E; hydro. energy is hydrophobic energy).
th(P,E,S,Ht).
In P,E, strand S has a total hydro. energy of Ht.
avh(P,E,S,Ha).
In P,E strand S has an average hydro. energy of Ha.
rth(P,E,S,Htr).
In P,E strand S has a rank of total hydro. energy of Htr.
rhy(P,E,S,Har).
In P,E strand S has a rank of average hydro. energy of Htr.

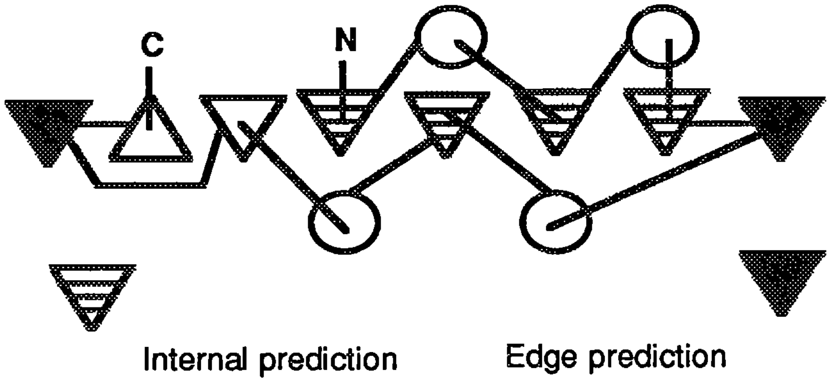

Figure 1. Schematic diagram showing edge and not_edge predictions for the protein Dihydrofolate reductase: circles are alpha-helices, triangles beta-sheets.

Each piece of secondary structure (strand, helix, turn, and loop) has a number that allows its position to be related in sequence to the other structures using arithmetic relational predicates. Typed arithmetical information is also given for the various predicates above. Typing prohibits spurious rules e.g. equating the rank of total hydrophobic energy to the number of helices separating it to the next strand (it is an inductive bias).

pairs1(P,S1,S2).
In protein P strand S2 follows strand S1.
pairs2(P,S1,S2).
In protein P strand S1 precedes strand S2.
(pairs1 and pairs2 have opposite GOLEM determinations)
pred(A,B).
B is A + 1.
succ(A,B).
A is A - 1.
(pred and succ have opposite GOLEM determinations)
lt(A,B).
A is less than B.

For each of the data sets there were 9263 background facts in the training data and 8047 background facts in the test data. The training data for the predicate edge consisted of 71 positive examples and 91 negative examples. Henceforth the construct [X : Y] is used as an abbreviation for "X positive examples and Y negative examples". The test data consisted of [27 : 42] examples. The training data for the predicate parallel consisted of [66 : 46] examples. The test data consisted of [39 : 11] examples. The training data for the predicate adjacency consisted of [127 : 786] examples. The test data consisted of [58 : 360] examples. The training data for the strand direction predicate consisted of [36 : 60] examples. The test data consisted of [18 : 26] examples.

Results

Rules for the Edge of a Sheet

Three rules were learnt for edge, and one rule for not_edge. These rules covered in the training examples all but [14 : 35] examples and all but [6 : 14] test examples. Accuracy is defined as the number of correctly predicted examples divided by the total number of predictions. Coverage is the number of correctly predicted examples divided by the total number of possible correct predictions. The coverage and accuracy given is that for each rule taken in isolation.

Rule edge1:
[Train cov. 0.465, acc 0.891: Test cov. 0.482, acc. 0.765]
edge(A,B) :- rth(A,C,B,rth0).

Rule edge2:
[Train cov. 0.310, acc. 0.880: Test cov. 0.148, acc. 0.500]
edge(A,B) :- pairs2(A,C,B), pairs2(A,D,C), seql(A,B,l3).

Rule edge3:
[Train cov. 0.197, acc. 1.000: Test cov. 0.296, acc. 1.000]
edge(A,B) :- seql(A,B,l2).

Rule not_edge1:
[Train cov. 0.550, acc. 0.909: Test cov. 0.4681, acc. 0.815]
not_edge(A,B) :-
pred(F,E), pred(H,G), pred(I,H), rhy(A,D,B,E),
rth(A,D,B,G), con_to_chain2(A,B,C).

The ordering of strands in sheets by their hydrophobic energy was first stressed by Sternberg and Thornton (Sternberg & Thornton 1977a) who noted that "the most hydrophobic strand is buried and the remaining strands are arranged in decreasing hydrophobicity outwards in both directions". Rule edge1, states that the strand of lowest rank hydrophobicity in a sheet is at an edge, this is implied by the Sternberg and Thornton statement. Sternberg and Thornton further noted that total hydrophobic energy was a better measure than average hydrophobic energy for predicting the ordering of strands. This is still true, and is reflected in GOLEM's choice of total hydrophobic energy. The best way found by GOLEM to characterise the remaining edge strands was to examine strand length, rules edge2 and edge3. Strand length is closely related to total hydrophobic energy. Two conditions of rule not_edge1 deal with hydrophobic energy: that the rank of total hydrophobic energy of an edge is at least two, and it does not have the lowest rank of average hydrophobic energy. These are consistent with the current understanding of protein structure. For a strand to be internal it should have a reasonable amount of not too diffuse hydrophobic energy. The other condition in not_edge1, that the following strand should be in the same sheet, is more difficult to understand. It may be that the connecting to a strand in another sheet disrupts hydrophobic ordering because of the packing of the edge to the other sheet.

Rules for Strand Orientation

Two rules were learnt for parallel, and one rule for not parallel. These rules covered in the training examples all but [7 : 8] examples and [8 : 2] test examples. The coverage and accuracy of each rule is that for each rule taken in isolation.

Rule parallel1:
[Train cov. 0.636, acc. 0.977: Test cov. 0.564, acc. 1.000]

```
parallel(A,B,C) :-
    pred(F,E), rhy(A,D,B,E),
    con_helix(A,B,C,G), pred(H,G).
```

Rule parallel2:
[Train cov. 0.500, acc. 0.868: Test cov. 0.410, acc. 0.833]

```
parallel(A,B,C) :-
    con_length(A,B,C,D), con_helix(A,B,C,F), succ(D,E),
    pred(G,F), rth(A,H,C,I), pred(J,I), seql(A,C,K),
    seql(A,B,L), succ(L,M), lt(K,M).
```

Rule not_parallel1:
[Train cov. 0.696, acc. 0.865: Test cov. 0.727, acc. 0.667]

```
not_parallel(A,B,C) :-
    pred(D,B), pred(E,D), con_helix(A,B,C,hi0).
```

Rule parallel1 has two conditions. The first condition, is that the strands are connected by at least 1 helix. This is consistent with current understanding of protein structure; if there was no helix, it is unlikely that connection between the strands would be long enough to allow the protein chain to loop back to the start of the first strand. The second condition, that strand B should not have the lowest rank of average hydrophobic energy in its sheet, suggests that parallel strands are more hydrophobic than anti-parallel strands. This idea is supported by the work of (Lifson & Sander 1979). The first condition of parallel2 is, that strand B should not have the lowest rank of average hydrophobic energy. The other conditions of parallel2 are: that sheets B and C are connected by at least 1 helix, and that the length of B is greater or equal to C. This final condition is interesting, it suggests that there is a relatively greater chance of being anti-parallel when the second strand is of greater length than the first. Such strands are relatively rare (46 examples) compared to pairs where the first strand if of greater length (71 examples). It is interesting to note how the relation A >= B is represented by the Prolog program induced by GOLEM. The relation is not given in the background knowledge, where there is only A > B; the relation is therefore expressed by A + 1 > B. Rule not-parallel1 is basically the negation of rules Parallel1 and Parallel2, its most important condition being that no helix separates the strands.

Rules for Strand Adjacency

Two rules were learnt for adjacent, and three rules for not adjacent. These rules covered in the training examples all but [60 : 234] examples and [18 : 89] test examples.

Rule adj1:
[Train cov. 0.441, acc. 0.849: Test cov. 0.345, acc. 0.833]

```
adj(A,B,C) :-
    con_length(A,B,C,D), succ(D,E), seql(A,C,F),
    succ(F,G), seql(A,B,H), pred(I,H), lt(I,G), succ(H,J),
    pred(K,F), lt(K,J).
```

Rule adj2:
[Train cov. 0.354, acc. 0.818: Test cov. 0.349, acc. 0.870]

```
adj(A,B,C) :-
    con_helix(A,B,C,D), pairs2(A,E,B), rth(A,F,B,G),
    pred(H,G), seql(A,C,I), seql(A,B,J), succ(J,K), lt(I,K).
```

Rule not_adj1:
[Train cov. 0.396, acc. 0.978: Test cov. 0.386, acc. 0.979]

```
not_adj(A,B,C) :-
    th(A,D,B,E), succ(E,F), succ(F,G), rhy(A,H,C,I),
    rhy(A,D,B,J),succ(J,K), lt(I,K), pairs1(A,B,L),
    pairs1(A,L,M), pred(N,C), pred(O,N), lt(M,O).
```

Rule not_adj2:
[Train cov. 0.313, acc. 0.984: Test cov. 0.397, acc. 0.960]

```
not_adj(A,B,C) :-
    avh(A,D,B,E), pairs2(A,F,B), avh(A,D,F,G),
    pairs2(A,H,C), lt(B,H), pred(I,F), pred(J,I),
    pairs1(A,B,K), rth(A,L,K,M), pred(N,M).
```

Rule not_adj3:
[Train cov. 0.350, acc. 0.984: Test cov. 0.419, acc. 0.968]

```
not_adj(A,B,C) :-
    pairs1(A,C,D), con_length(A,C,D,E), succ(E,F),
    th(A,G,B,H), succ(H,I), succ(I,J),
    con_to_chain1(A,C,K), pred(L,K), pred(M,L),
    succ(B,N), pred(O,C), pred(P,O), lt(N,P).
```

The most important feature that was found in the rules for predicting adjacency, is that the two strands must follow in sequence (Richardson 1981). This illustrates the importance of short range interactions in sheet formation. Note, however, that in the context of packing prediction, short range means a far greater distance than it would in predicting secondary structure. The reason that sequentially adjacent strands are likely to pack adjacently in space is due to them having a far greater chance of interacting during the folding process. Rule adj1 has the interesting condition that the stands must differ in length by less than 2 residues. This is consistent with the current understanding of protein structure, as adjacent strands should be about the same length, so that they can form hydrogen bonds. Note how this condition is represented in GOLEM, it is built up from the greater-than relation (<) and the successor relation (+1) in the following way: ((((A + 1) + 1) > B) and (((B + 1) + 1) > A))). The discovery of such a complicated relation between two length illustrates the power of the ILP methodology. A conventional attribute based learning system would not have great difficulty in learning this concept.

The rules for predicting that two strands will not be adjacent in a sheet are the most complicated found in this study. This reflects the larger number of examples of non-adjacency compared to adjacency, and the need for highly accurate rules. The most important condition preventing strands becoming adjacent is that the strands are separated by a sufficient distance. This is the reverse of condition of the most important condition for strands being adjacent.

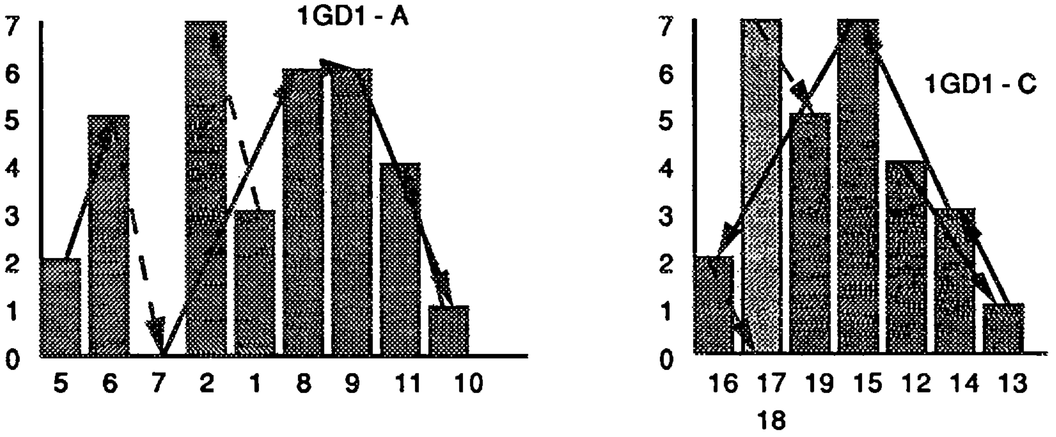

Figure 2. Predictions of winding direction. The Y-axis is the rank of total hydrophobic energy. The X-axis is the sequential number of strand, i.e. strand no. x is the nth in sequence. Arrows show the prediction of winding direction; dotted arrows are incorrect predictions. In sheet 1GD1-C, strands 17 and 18 have the same position in the sheet - causing two prediction errors.

Rules for Strand Direction

One rule were learnt for in, and one rule for out (not in). These rules covered in the training examples all but [10 : 11] examples and [2 : 3] test examples. The coverage and accuracy of each rule is that for each rule taken in isolation.

Rule in1:
[Train cov. 0.500, acc. 0.818: Test cov. 0.556, acc. 0.833]
in(A,B,C) :-
con_to_chain1(A,B,D), con_length(A,B,C,E), succ(E,F), rth(A,G,B,H), rth(A,G,C,I), pred(J,I), lt(H,J).

Rule out1:
[Train cov. 0.750, acc. 0.849: Test cov. 0.808, acc. 0.777]
out(A,B,C) :-
con_to_chain2(A,B,D), succ(D,E), succ(E,F), rth(A,G,C,H), rth(A,G,B,I), succ(I,J), lt(H,J).

In rule in1 the most important condition is that the rank total hydrophobic energy of C should be 2 ranks greater than B. This is consistent with the analysis of Sternberg and Thornton (Sternberg & Thornton 1977a); who proposed that strands are ordered in terms of their hydrophobicities, with the most hydrophobic strands in the centre. In rule out1 the most important condition is that the total hydrophobic energy of B greater than or equal to C. This is almost the reverse of rule in1, although it is interesting that the condition for moving in is more strict than moving out, i.e. it is easier to wind out. To illustrate the application of the rules learnt by GOLEM the results of applying rules in1 and out1 to glyceraldehyde dehydrogenase are shown in Fig. 2.

It is important to note that there are more connections in the direction towards the edge than there are away from the edge (significant at 1% level); this is illustrated in Fig. 3. This figure also neatly illustrates why the discrimination between rules in1 and out1 in terms of difference in total hydrophobic energy is placed where it is. Where the difference in ranks is two or greater, then there are more examples of in compared to out: where the difference is zero or less there are more examples of out: the case is ambiguous where the difference is 1.

Discussion

Winding Direction

A number of the rules found by GOLEM are thought to be consequences of the fact that beta-sheets in alpha-beta proteins are more likely to start winding from the centre out, than from the edge in. This tendency has been hinted at in the literature (Richardson 1981), but the logical consequences of it have not been fully appreciated. The consequences were discovered by consideration of the rules generated by GOLEM. One consequence is that because proteins are roughly globular in shape, the relationship between pairs of sequential strands is one of decreasing size, i.e. the first strand is likely to be larger than the second. This constraint may be usefully applied back to secondary structure prediction. Winding from the centre out also explains why there are more pairs of strands in the direction out than in (Fig. 3). It takes a relatively large change in rank total hydrophobic energy between two sequential sheets to change the direction of winding.

Comparison with Hand Crafted Rules

A number of hand crafted rules have been proposed for constraints in the folding of beta strands in alpha/beta proteins, see discussion in Clark *et al.* (1991). It is interesting to compare these rules with those found by GOLEM :

- Some rules were found to be identical, e.g. the rule: "the least hydrophobic strand is on the edge of the sheet (using total hydropathies)", is identical in meaning with rule edge1. This rule originated in the work of Sternberg and Thornton (Sternberg & Thornton 1977a).
- Other rules are closely related but are somewhat more general, e.g. parallel beta-alpha-beta connections should contain at least 10 residues in the coil (Taylor & Green 1989).
- However, some of the proposed rules for nucleotide binding domains did not generalise to all alpha/beta proteins. For example, Branden (1980) proposed the rule: that the initial strand in sequence is not an edge sheet. This can be translated into Prolog as:
  [Train cov. 0.127, acc. 0.600 : Test 0.185, 0.833]
  not_edge(A,S) :- lt(1,S).
  This rule has rather a low accuracy and coverage compared to the rules found by GOLEM.
- Other proposed rules could not have been found because of the limitation of the representation used. For example, Taylor and Green (Taylor & Green 1989) proposed that unconserved strands are at the edge. This rule could not have found as the concept of conservation of sequence is not represented in the background knowledge; it relates to the rate of mutation across homologous proteins. Similarly, Richardson (1981) proposed that there is only one change in winding direction in a sheet. Such a rule could not have been found because the concept of change in winding direction is second order compared to strand direction.

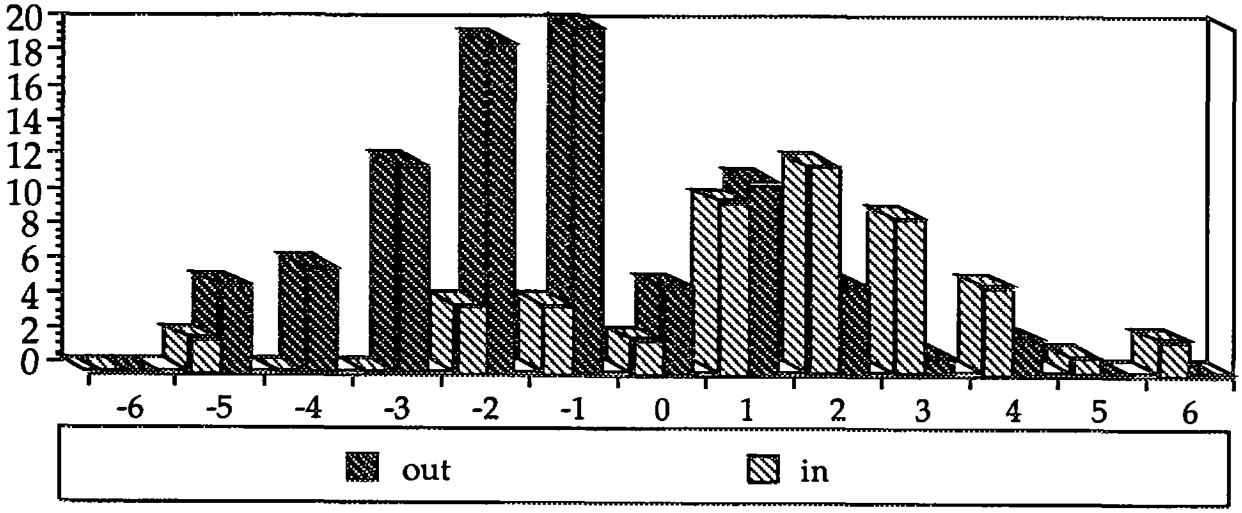

Figure 3. The difference in rank of total hydrophobic energy by winding direction. In means that the strand winds towards the centre of the sheet: out means that strand winds towards the edge of the sheet. The X-axis is the difference in rank of total hydrophobic energy. The Y-axis is the number of strand pairs with the difference in rank of total hydrophobic energy.

ElipSys Rules

The existing hand-produced rules concerned relating to the secondary structure packing of alpha/beta proteins have been gathered together by Clark *et al* (1991, 1993). These rules were formalised and translated into logic programs and used as part of an overall structural packing program. This program inputs one or more protein secondary structure assignments and outputs a predicted packing for the protein. Currently the rules and program are implemented in the parallel constraint logic programming language ElipSys. The success of these hand-produced rules at predicting protein packing patterns is currently being evaluated by a systematic prediction of the database. The rules discovered by GOLEM are being translated into ElipSys. It is intended to directly compare the success of these rules with the hand-produced ones on the systematic prediction of the database.

Machine Learning as a Tool for Molecular Biology

In this paper, machine learning is viewed as a tool to aid scientists in the discovery of patterns in data from molecular biology. It is considered to be an alternative method to: examining data "by eye" (perhaps using sophisticated visualisation software), or the use of statistical methods. We see the process of applying machine learning to a scientific problem as an interactive cycle between the problem formulation by the scientist and application of a machine learning program.

- The scientist first identifies the problem of interest.
- A formal data representation is created that is intended to capture the important features of the problem.
- The actual data is then translated into this formalism to produce a model of the problem.
- This data is input into the machine learning system and rules generated.
- The scientist then examines the rules to gleam any regularities and patterns in the data that are of scientific interest.
- If this cycle is successful the newly discovered regularities can be used to form a better data representation and the cycle repeated.

Conclusions

This paper tests the application of ILP to the discovery of patterns in data in molecular biology. ILP was used to discover constraints in the packing of beta-sheets in alpha/beta proteins. Constraints (rules) were learnt for four features of beta-sheet packing: the winding direction of two sequential strands (two rules found), whether two sequential strands pack parallel or anti-parallel (three rules found), whether two strands pack adjacently (five rules found), and whether a beta-strand is at an edge (four rules found). Investigation of the rules revealed interesting patterns, some of which were previously known, others are that are novel. Novel features include two previously unrecognised consequences of the bias of winding direction to start from the centre: that the relationship between pairs of sequential strands is in general one of decreasing size, i.e. the first strand is likely to be larger than the second; and that more sequential pairs of strands wind in the direction out than in.

Acknowledgements

We would like to thank Chris Rawlings and Steve Gardner (Oxford Molecular) for making IDITIS raw data available, and Steve Muggleton for supplying GOLEM.

References

Branden, C. (1980). *Q. Rev. Biophys* **13**: 317-338.

Clark, D. A., Rawlings, C.J., Shirazi, J., Veron, A. and Reeve, M. (1993). Protein topology prediction through parallel constraint logic programming. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, AAAI/MIT Press, Menlo Park.

Clark, D. A., Shirazi, J. and Rawlings, C.J. (1991). Protein topology prediction through constraint-based search and the evaluation of topological folding rules. *Prot. Engng.* **4**: 751-760.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**: 97-120.

Islam, S. A. and Sternberg M.J.E. (1989). A relational database of protein structures designed for flexible enquires about conformation. *Prot. Engng.* 2(6): 431-442.

King, R. D., Muggleton, S., Lewis, R. and Sternberg, M.J.E. (1992). Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. Natl. Acad. Sci. USA* **89**: 11322-11326.

King, R.D., Clark, D.A., Shirazi, J., Sternberg, M.J.E. (1994) Discovery of protein structural constraints in a deductive database using inductive logic programming. In *Machine Intelligence 14* eds. D. Michie, S. Muggleton, K. Furakawa. Oxford University Press. Oxford. (in press)

Lifson, S. and Sander, C. (1979). Antiparallel and parallel beta-strands differ in amino acid residue preferences. *Nature* **282**. 109-110.

Muggleton, S. and Feng C. (1990). Efficient induction of logic programs. In *Proceedings of the First Conference on Algorithmic Learning Theory,* Tokyo, Jpn. Soc. Artificial Intelligence.

Muggleton, S., King, R.D. and Sternberg, M.J.E. (1992). Protein secondary structure prediction using logic-based machine learning. *Protein Engng.* 5(7): 647-657.

Orengo, C. A., Flores, T.P., Taylor, W.R. and Thornton, J.M. (1993). Identification and classification of protein fold families. *Prot. Engng.* 6(5): 485-500.

Rawlings, C. J. E., Taylor, W.R., Nyakairu, J. Fox, J. and Sternberg, M.J.E. (1985). TOPOL. *J. Mol. Graphics* 3(4): 151-157.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry* **34**: 167-339.

Rost, B. and Sander C. (1993). Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J. Mol. Biol.* **232**: 584-599.

Schulz, G. E. and Schirmer R.H. (1978). *Principles of Protein Structure*. Springer-Verlag.

Shirazi, J. and Clark, D.A. (1993). Combining Databases and constraints in protein structural analysis. Technical report ESPRIT project 6708 "Applause".

Sternberg, M. J. E. and Thornton J.M, (1977a). On the conformation of proteins: hydrophobic ordering of strands in beta-pleated sheets. *J. Mol. Biol.* **115**: 1-17.

Sternberg, M. J. E. and Thornton J.M. (1977b). On the conformation of proteins: towards the prediction of strand arrangements in beta-pleated sheets. J. M. Biol. **113**: 401-418.

Taylor, W. R. and Green N.M. (1989). The predicted secondary structure of the nucleotide-binding sites of six cation-transporting ATPases leads to a probable tertiary fold. *European Journal of Biochemistry* **179**: 241-248.