PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Application of Machine Learning to Structural Molecular Biology

Michael J. E. Sternberg, Ross D. King, Richard A. Lewis and Stephen Muggleton

Phil. Trans. R. Soc. Lond. B 1994 **344**, 365-371
doi: 10.1098/rstb.1994.0075

Application of machine learning to structural molecular biology

MICHAEL J. E. STERNBERG[1], ROSS D. KING[1,2], RICHARD A. LEWIS[1]§ AND STEPHEN MUGGLETON[2]‡

[1]*Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX, U.K.*
[2]*Turing Institute, George House, 36 North Hanover Street, Glasgow G1 2AD, U.K.*

SUMMARY

A technique of machine learning, inductive logic programming implemented in the program GOLEM, has been applied to three problems in structural molecular biology. These problems are: the prediction of protein secondary structure; the identification of rules governing the arrangement of β-sheets strands in the tertiary folding of proteins; and the modelling of a quantitative structure activity relationship (QSAR) of a series of drugs. For secondary structure prediction and the QSAR, GOLEM yielded predictions comparable with contemporary approaches including neural networks. Rules for β-strand arrangement are derived and it is planned to contrast their accuracy with those obtained by human inspection. In all three studies GOLEM discovered rules that provided insight into the stereochemistry of the system. We conclude machine leaning used together with human intervention will provide a powerful tool to discover patterns in biological sequences and structures.

1. INTRODUCTION

The developments of gene cloning and sequencing are leading to an explosion of information in molecular biology. The identification of a new protein sequence via its gene raises the question of characterizing its three-dimensional structure. As experimental structure determination by crystallography and nuclear magnetic resonance remain time-consuming and dependent on milligrams of material, the theoretical approach of predicting protein conformation from sequence is of increased importance. Additionally, characterization of the function of a protein leads to the question of designing molecules that can regulate its activity and may serve as a therapeutic agent. Computer modelling is widely used to quantify structure-activity relationships (QSAR) as a guide for drug design. Many theoretical approaches to these problems are empirical: rules are gleaned from observations.

In this paper we test the idea of applying machine learning as a tool to aid scientists in the discovery of patterns in biological data. Machine learning is considered to be an alternative method to visual examination of data (perhaps aided by sophisticated graphics), or the use of statistical methods. In particular we use inductive logic programming (ILP) as implemented in the computer program GOLEM (Muggleton & Feng 1990), to derive rules that relate protein chemical structure, including sequence, into information about conformation and function. Three application areas are considered: (i) the prediction of protein secondary structure from sequence; (ii) the identification of rules describing the three-dimensional folding of β-sheet strands in globular proteins as a step towards tertiary structure prediction; and (iii) the derivation of a QSAR for a series of drugs that bind to a protein.

This work has broader implications for development of machine learning (Weiss & Kulikowski 1991) in addition to the specific advances in modelling in the particular areas. The application of machine learning to the discovery of patterns in scientific data is known as 'scientific discovery'. It is only by the application of machine learning to current scientific problems that the general field will advance. Molecular biology is an ideal test bed for applying machine learning to science: there are a number of suitable important problems, the increasing pace of data acquisition is swamping traditional methods of pattern discovery, the data is mostly in discrete symbolic form.

2. INDUCTIVE LOGIC PROGRAMMING BY GOLEM

In ILP a set of examples is examined and rules are derived which are expressed as logical relationships between objects. GOLEM (Muggleton & Feng 1990) encodes relationships in the first order predicate calculus which is sufficiently expressive to encode most of the concepts used to describe biological

§Present address: Discovery Chemistry Group, Rhône-Poulenc Rorer Ltd, Dagenham Research Centre, Rainham Road South, Dagenham RM10 7XS, U.K.
‡Present address: Oxford University Computing Laboratory, Wolfson Building, Parks Rd, Oxford OX1 3QD, U.K.

Phil. Trans. R. Soc. Lond. B (1994) **344,** 365–371
Printed in Great Britain

molecules. The aim is not simply to obtain predictive rules but also to derive understandable logical relationships. This approach contrasts with numerically based methods, including neural networks (Simpson 1990), that often have the drawback that the resultant rules are difficult to interpret. Technically the computer program, GOLEM, is written in C with the input and output of logical concepts encoded in the logic programming language PROLOG (Clocksin & Mellish 1981).

Figure 1 shows the general learning method used by GOLEM. The input consists of the observations encoded as positive and as negative examples together with the background knowledge describing the system. The learning process by GOLEM is:

1. Take two positive examples at random.
2. Generate a rule from the common properties of the examples.
3. Evaluate the accuracy of this rule on the remaining examples.
4. Repeat steps 1 to 3 several times (typically 10) and generate the most accurate rule.
5. Add an additional positive example and repeat the above to generate a more general rule.
6. Continue with step 5 until maximum coverage of examples by rule occurs, then store rule in background knowledge.
7. Start at step 1 to search for next rule.

3. SECONDARY STRUCTURE PREDICTION

In protein secondary structure prediction, the local sequence is examined to predict its main-chain conformation, in particular whether it adopts an α-helix, β-sheet strand or coil. One approach is to develop algorithms for a particular structural class of proteins and here we report the application to the all-α proteins that have α-helices and coil but little or no β-sheet (for details see Muggleton *et al.* (1992)).

A training set of 12 proteins was used. The input consisted of the observations of the location of α-helical residues in the protein. For example:

alpha(155C,110)

defines that residue 110 in the protein with code 155C is in an α-helix. The background information is of two types. The first defines the chemical structure of residues, for example that 110 in 155C is valine(V):

position(155C,110,V).

The second defines the chemical properties of the residues, for example that valine is hydrophobic:

hydrophobic(V).

GOLEM induced rules governing α-helix formation. However this lead to a speckled prediction with isolated residues being predicted as helical. Two cycles of smoothing were learnt by GOLEM.

The GOLEM rules yielded an accuracy of 78% for the number of residues correctly predicted α-helix or coil in the training set and of 81% in the test set. These values have an estimated error of ±2%. These results are better than a prediction by neural network on this α/α class that yielded 76% (Kneller *et al.* 1990). α-Helices have a periodicity of 3.6 residues per turn and the location of sequential residues can be plotted on a

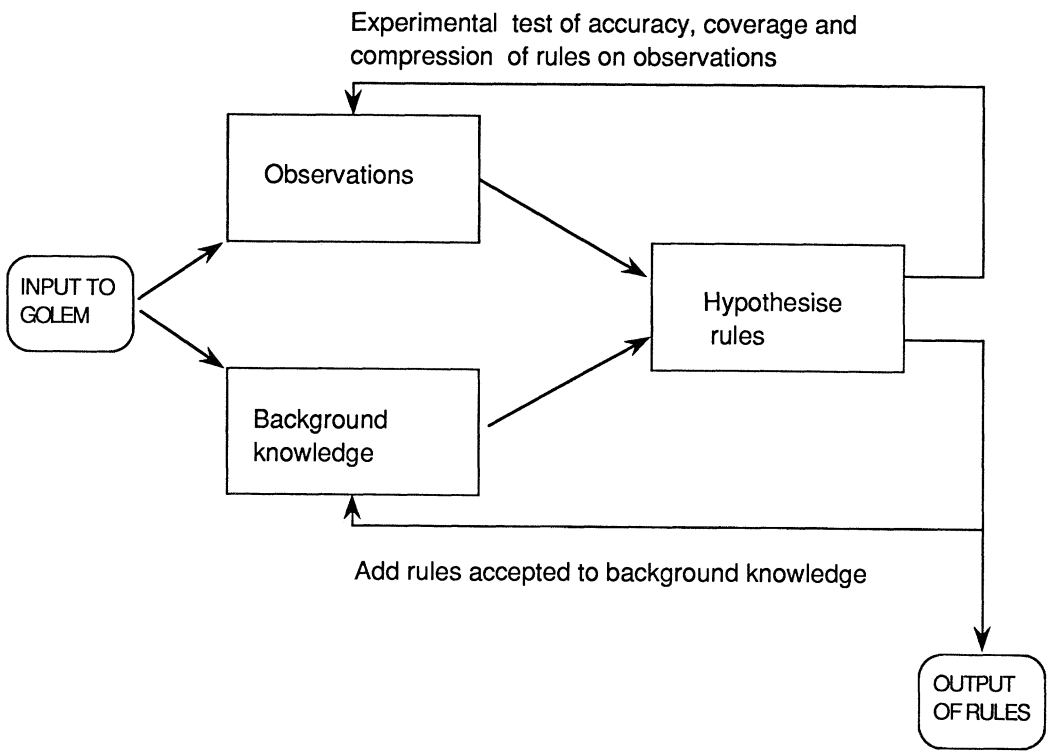

Figure 1. Inductive logic programming as implemented by GOLEM.

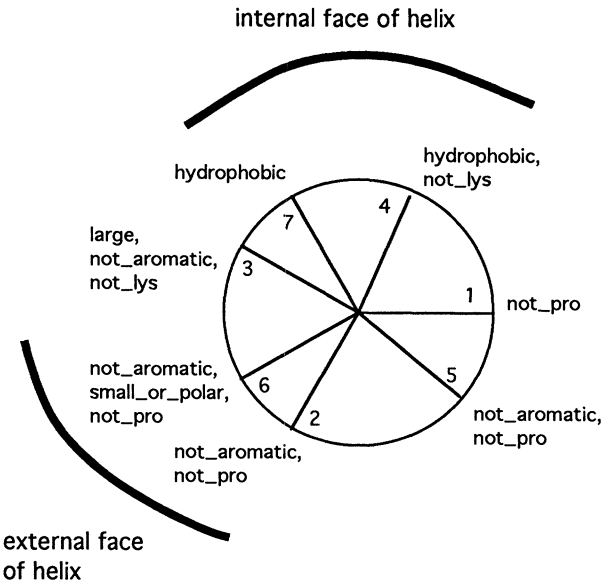

Figure 2. One rule for α-helical prediction mapped onto an α-helical wheel. The residue numbers are given together with the allowed amino acids at each position.

circle. Figure 2 plots one of the rules on such a helical wheel. GOLEM learnt the tendency for α-helices to have one face that is formed from hydrophobic residues and another more polar face. However, some of the more specific features of the rules such as the restriction that proline is only allowed at positions 3 and 7 on one face of the helix is not understood at present. Thus this rules suggests a new analysis of protein structure to identify the importance of this feature.

4. β-SHEET TOPOLOGY

One approach in the prediction of protein structure is first to identify the local secondary structure and then to determine the approximate three-dimensional packing of these α-helices and β-strands in the tertiary fold. Here we consider a subset of the problem: the folding of α/β proteins. In an α/β domain there tends to be an alternation of α-helices and β-sheet strands as one progresses along the chain. The fold is often represented in a schematic diagram (see figure 3). A key aspect of the fold is the arrangement of strands in the sheet defined by order and direction. Previous analyses of sheet arrangements have identified several principles (e.g. Richardson 1977, 1981; Sternberg & Thornton 1977*a*,*b*). Here we show that these rules and some new rules can be learnt automatically by GOLEM.

The data of β-sheet arrangements was taken from the PAPAIN database (Clark *et al.* 1991; Rawlings *et al.* 1985) that encodes structural features in PROLOG. A set of 23 non-homologous α/β structures from Orengo *et al.* (1993) was used and from these seven were randomly chosen as a test set. The object was to learn rules for the following features:

edge(P,S)

which states that strand S in protein P is positioned at the edge of a sheet.

in(P,S1,S2)

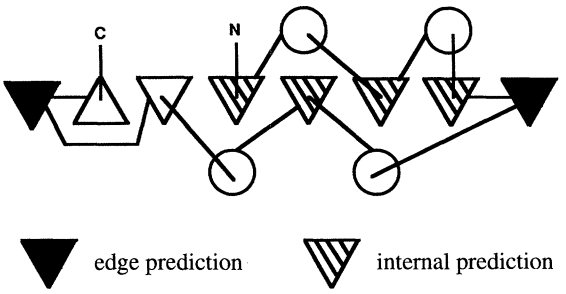

Figure 3. A schematic diagram of the arrangement of β-strands and α-helices in dihydrofolate reductase. Each β-strand is drawn as a triangle whose apex points up or down depending on whether it is viewed from the amino or carboxyl terminus. α-Helices are drawn as triangles. N and C denote the termini. The figure shows the prediction of edge and internal strands based on rules for edge and not_edge induced by GOLEM.

which formalizes the winding direction and states that strand S1 is closer to the edge than sequential strand S2 in protein P.

adj(P,S1,S2)

which states that sequential strands S1 and S2 in protein P are adjacent in a sheet.

parallel(P,S1,S2)

which states that sequential strands S1 and S2 in protein P are parallel.

The background knowledge encoded: the length of the strand together with several aspects about the nature of the connection between two strands including the length of the connection and whether this included an α-helix. There was also information about hydrophobicity of the strands evaluated as the free energy change (Kyte & Doolittle 1982) on transfering the residues in the strand from a non polar environment into water. Thus the more hydrophobic strands have a larger hydrophobic energy. The features encoded were: the total hydrophobic energy, the average hydrophobic energy per strand residue and the ranks of total and the average hydrophobic energies of the strand.

The rules learnt for edge, not_edge, in and not_in (out) will be used to illustrate the results. Three rules were learnt for edge, and one rule for not_edge. The simplest of the rules found for an edge strand expressed in PROLOG is:

edge(A,B): rth(A,C,B,rth0).

In English this reads:

a strand is at the edge if
    it has the lowest rank of total hydrophobic energy in its sheet.

On the testing set this covered 0.482 of edge strands and predicted with an accuracy of 0.765.

The rule for not_edge is:

a strand is not at the edge if
    it does not have the lowest rank of average hydrophobic energy in its sheet

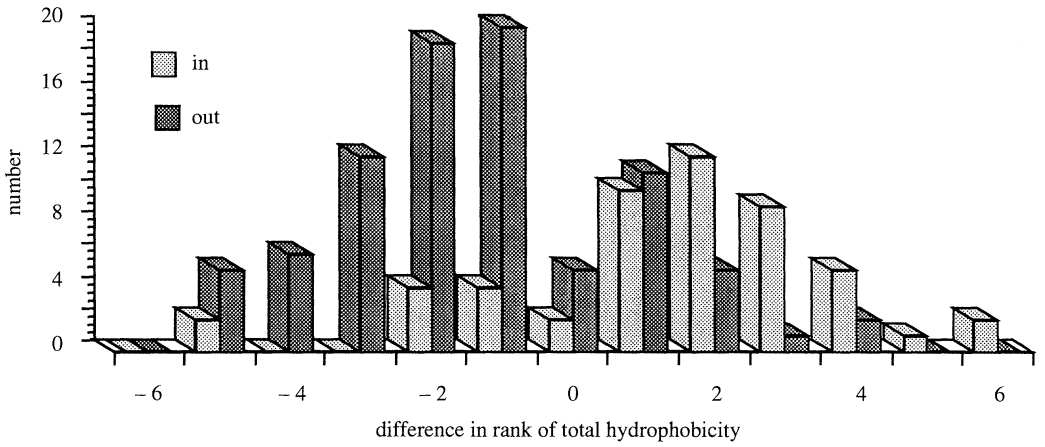

Figure 4. The difference in rank of total hydrophobicity of strands classified by winding direction. In denotes the second sequential strand is closer towards the centre of the sheet, out that the second strand goes towards the edge of the sheet. The figure shows that 'in' strands dominate only when the difference in rank is 2 or more.

and its rank of total hydrophobic energy in its sheet is at least 2 and
the following strand is in the same sheet.

One rule was learnt for in:

strand B is closer to the edge of the sheet than strand C when C directly follows B in sequence if
the rank of total hydrophobic energy of C is 2 greater than B and
they are connected <50 residues.

One rule for out (not_in):

strand C is closer to the edge of the sheet than strand B when C directly follows B in sequence if
the rank of total hydrophobic energy of B $\geqslant$C and
B is not the last strand in the sheet.

These rules agree and formalize an earlier analysis that suggested strands are ordered in terms of their hydrophobicities, with the most hydrophobic strands in the centre (Sternberg & Thornton 1977*a*). Importantly they also suggest new features that might well remain unnoticed by simple human examination. For example, figure 4 illustrates the rule that for a pair of sequential strands to progress into a sheet requires an increase in rank of total hydrophobicity of two but going out requires the same or lower rank. It is planned to explore the extent to which these rules can be used to aid in protein structure prediction. The in–out rules promise to be effective. Figure 5 shows the prediction of winding direction on a four sheets in the test set. On all the test data, 31 directions were correctly predicted, eight incorrectly and only five sequential strands were not covered by the rule.

5. DRUG DESIGN

The modelling of a quantitative structure activity relationship (QSAR) for a series of compounds remains

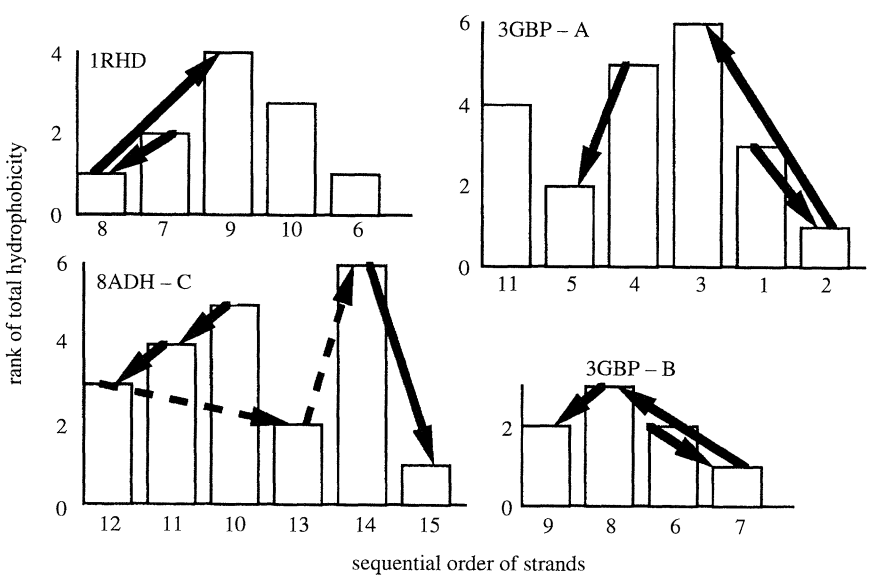

Figure 5. Prediction of winding direction of strands. Solid arrows correct predictions for two sequential strands, broken arrows incorrect predictions. The sheets are: 1RHD in rhodanese; 8ADH-C in liver alcohol dehydrogenase; 3GBP-A and 3GBP-B in galactose binding protein.

a central tool in the systematic design of drugs. The object is to relate the chemical properties of substituents on a core pharmacaphore to the activity of the molecules. The traditional approach stems from the pioneering work of Hansch and coworkers (Hansch *et al.* 1962, 1982; Hansch 1969) and effectively employs nonlinear regression from descriptors of the properties of the substituents. The most commonly used descriptors are related to hydrophobicity and volume. Recently neural networks have been applied to derive a QSAR taking the Hansch parameters as input (see, for example, Andrea & Kalayeh 1991) and reported marked improvements over linear regression. We have, however, repeated this study with cross validation trials in which the test data is chosen at random rather than matched to the training data. Our study finds no statistically significant improvement by neural networks over nonlinear regression (Hirst *et al.* 1994*a*,*b*; King *et al.* 1993).

A major problem with both the Hansch approach and neural networks is that they provide little insight into the stereochemistry of the drug–receptor interaction that can guide the design of new compounds. ILP, which reasons symbolically rather than numerically , provides an approach to glean these insights.

GOLEM was applied (King *et al.* 1992; Hirst *et al.* 1994*a*) to model the inhibition of *E. coli* dihydrofolate reductase by 2,4-diamino-5-(substituted-benzyl) pyrimidines (Hansch *et al.* 1982; Roth *et al.* 1987; Selassie *et al.* 1991) as exemplified by the drug trimethoprim (figure 6*a*). The activity of the compounds provided the observations. The positive examples were of the form:

great(drug55,drug9)

stating that drug 55 has greater inhibition than drug 9. The negative examples were false statements and were simply the converse:

great(drug9,drug55).

The background knowledge defined the chemical structure of the drugs, e.g.

streak(drug55,Cl,NH_2,CH_3)

which defines the substitutions at positions 3, 4 and 5. In addition, the properties of each substituent is expressed by descriptors, called physico-chemical attributes, which are designed to help in the formation of human understandable rules. The attributes are: polarity, size, flexibility, hydrogen-bond donor, hydrogen-bond acceptor, π donor, π acceptor, polarizability, branching and effect. This was represented using different predicates for each property and value, e.g.

polar(Br,polar3)

states that Br has polarity of value 3.

Five trials of the approach were performed corresponding to random, non overlapping splits of the data of the 55 drugs into testing sets of 11 drugs with the remaining 44 forming the training set (Hirst *et al.* 1994*a*). Over these five trials the average Spearman rank correlation for the agreement between predicted and true rank of the testing drugs was 0.68 ± 0.11. To represent the traditional Hansch approach while avoiding any bias due to the use of different representations, the same data was modelled by a regression on linear and squared terms of the attributes. The average Spearman rank was 0.65 ± 0.10. The better performance of GOLEM is not statistically significant.

A major aim in this work was to obtain insight into the stereochemistry of drug–receptor interactions. An example of a rule generated is:

drug A is better than drug B if:
  drug A has a substituent at position 3 with
    hydrogen-bond donor = 0 and
    π-acceptor = 0 and
    polarity > 0 and
    size < 3 and
  drug A has a substituent at position 4 (i.e. not hydrogen) and
drug B has no substituent at position 5.

In total 59 rules were found for the five cross-validation runs on the pyrimidines. From these rules, seven consensus rules were formed manually by selecting the most commonly found features These consensus rules (not shown) are simpler in form and easier to understand than the automatically generated rules. The consensus rules were tested against the five

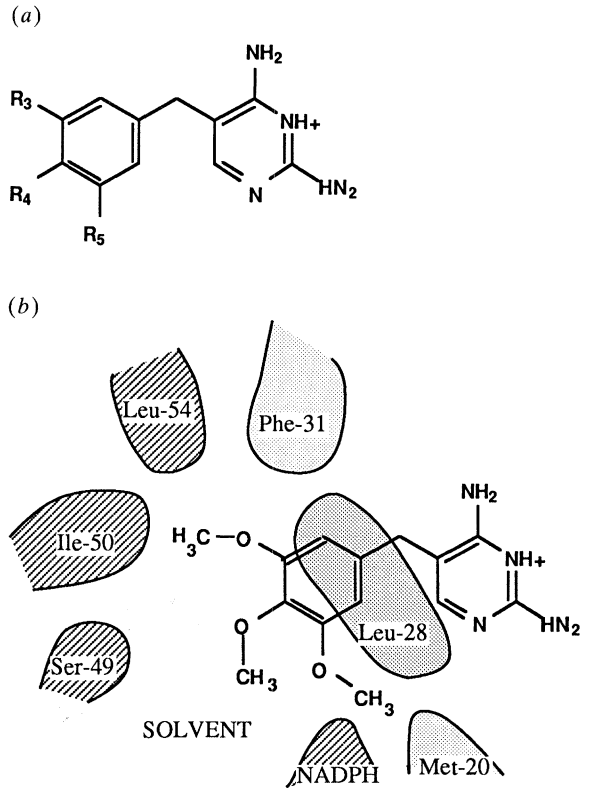

Figure 6. Trimethoprim binding to dihydrofolate reductase. (*a*) Structure of trimethoprim analogues; (*b*) a cartoon of the interaction of trimethoprim with dihydrofolate reductase from the X-ray structure (Champness *et al.* 1986), faint stippling indicates that the residue lies below the plane of the phenyl ring, diagonal lines that the atoms are above.

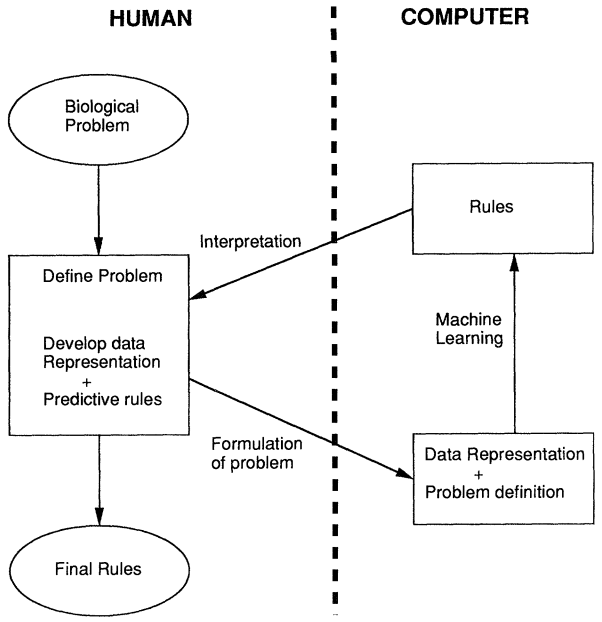

Figure 7. Summary of the approach envisaged for the application of machine learning to scientific problems.

cross-validation data sets, giving an average Spearman rank correlation on the test data of 0.845. It is straightforward to use the consensus rules to generate the best predicted drug or drugs. Considering positions 3 and 5, the only possible substituents, using all physico-chemical attributes, with the conjunction of: polarity <5, size <3, hydrogen-bond donor $=0$, π-donor $=0$, σ-effect <5, flexibility <3, are OCH3, I, Cl, and Br (O excluded). These are therefore the substituents recommended for positions 3 and 5; the rules do not distinguish between these groups. There are far fewer constraints at position 4, only the conjunction of: polarity <5, and hydrogen-bond donor $=0$.

Figure 6*b* is a cartoon of the stereochemistry of trimethoprim binding to *E. coli* dihydrofolate reductase as revealed at atomic resolution by protein crystallography (Champness *et al.* 1986). The 3 position is not exposed to solvent which is accord with features of the suggestions of restraints on its structure including that it should not be a hydrogen bond donor at this position. The 4 position is more exposed to solvent and has fewer constraints.

6. CONCLUSION

Figure 7 summarizes the approach envisaged for the application of machine learning to scientific problems. There is an interactive cycle between human analysis and machine leaning. Initially traditional methods process the data and develop representations that characterize the system and rules describing the relationship between the components of the system. Next machine learning uses these representation to identify new, and hopefully more powerful and incisive, rules. Then human intervention is required for interpretation of the rules and the cycle can be repeated.

The machine learning approach encoded in GOLEM has been consistently applied to three distinct applications in structural molecular biology. The predictions made for identifying α-helical secondary structures and the rank order of drug activity are at least as good as statistical approaches including neural networks. In both applications, rules were generated that encoded stereochemical principles. In drug design, these principles were in agreement with the knowledge of the receptor (dihydrofolate reductase) revealed by protein crystallography. Thus machine learning was extracting information that might well direct a drug discovery programme. In the work on the arrangements of β-sheet, GOLEM rediscovered the principle of hydrophobic ordering. In the three areas there were common features in representing the problem into a form suitable for machine learning as implemented by GOLEM. Chemical structure such as hydrophobicity was assigned explicitly to the objects be they residues, drug substituents or β-strands. The resultant rules reasoned based on these chemical attributes. We conclude that machine learning exemplified in GOLEM provides a powerful approach to extract principles from ever-increasing wealth of biological data.

We thank Dr Dominic Clark, Dr Christopher J. Rawlings and Dr Jack Shirazi for the PAPAIN database of protein structural features; Dr Jonathan Hirst and Dr Ashwin Srinivasan for helpful discussions and Professor Donald Michie for encouragment in this work.

REFERENCES

Andrea, T.A. & Kalayeh, H. 1991 Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *J. med. Chem.* **34**, 2824–2836.

Champness, J.N., Stammers, D.K. & Beddell, C.R. 1986 Crystallographic investigation of the cooperative interaction between trimethoprim, reduced cofactor and dihydrofolate reductase. *FEBS Lett.* **199**, 61–67.

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

Clocksin, W.F. & Mellish, C.S. 1981 *Programming in Prolog*. Berlin: Springer-Verlag.

Hansch, C. 1969 A quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.* **2**, 232–239.

Hansch, C., Li, R.-I., Blaney, J.M. & Langridge, R. 1982 Comparison of the inhibition of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase by 2,4-diamino-5-(substituted-benzyl) pyrimidines: quantitative structure-activity relationships, X-ray crystallography, and computer graphics in structure-activity analysis. *J. med. Chem.* **25**, 777–784.

Hansch, C., Maloney, P.P., Fujita, T. & Muir, R.M. 1962 Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature, Lond.* **194**, 178–180.

Hirst, J.D., King, R.D. & Sternberg, M.J.E. 1994*a* Quantitative structure-activity relationships: neural networks and inductive logic programming compared against statistical methods I. The inhibition of dihydrofolate reductase by pyrimidines. *J. comput. aided molec. Des.* (In the press.)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Hirst, J.D., King, R.D. & Sternberg, M.J.E. 1994*b* Quantitative structure-activity relationships: neural networks and inductive logic programming compared against statistical methods II. The inhibition of dihydrofolate reductase by triazines. *J. comput. aided molec. Des.* (In the press.)

King, R.D., Hirst, J.D. & Sternberg, M.J.E. 1993 New approaches to QSAR: neural networks and machine learning. *Persp. Drug Discov. Design* (In the press.)

King, R.D., Muggleton, S., Lewis, R.A. & Sternberg, M.J.E. 1992 Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationship of trimethoprim analogues binding to dihydrofolate reductase. *Proc. natn. Acad. Sci. U.S.A.* **89**, 11322–11326.

Kneller, D.G., Cohen, F.E. & Langridge, R. 1990 Improvements in protein secondary structure prediction by an enhanced neural network. *J. molec. Biol.* **214**, 171–182.

Kyte, J. & Doolittle, R.F. 1982 A simple method for displaying the hydropathic character of a protein. *J. molec. Biol.* **157**, 105–132.

Muggleton, S. & Feng, C. 1990 Efficient induction of logic programs. In *Proceedings of the first conference on algorithmic learning theory* (ed. S. Arikawa, S. Goto, S. Ohsuga & T. Yokomori), pp. 368–381. Tokyo: Japanese Society for Artificial Intelligence.

Muggleton, S., King, R.D. & Sternberg, M.J.E. 1992 Protein secondary structure prediction using logic. *Prot. Eng.* **5**, 647–657.

Orengo, C.A., Flores, T.P., Taylor, W.R. & Thornton, J.M. 1993 Identification and classification of protein fold families. *Prot. Eng.* **6**, 485–500.

Rawlings, C.J., Taylor, W.R., Nyakairu, J., Fox, J. & Sternberg, M.J.E. 1985 Reasoning about protein topology using the logic programming language PROLOG. *J molec. Graph.* **151**, 151–157.

Richardson, J.S. 1977 β-Sheet topology and the relatedness of proteins. *Nature, Lond.* **268**, 495–500.

Richardson, J.S. 1981 The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**, 167–339.

Roth, B., Rauckman, B.S., Ferone, R., Baccanari, D.P., Champness, J.N. & Hyde, R.M. 1987 2,4-diamino-5-benzylpyrimidines as antibacterial agents. 7. Analysis of the effect of 3,5-dialkyl substitutent size and shape on binding to four different dihydrofolate reductase enzymes. *J. med. Chem.* **30**, 348–356.

Selassie, C.D., Li, R.-L., Poe, M. & Hansch, C. 1991 On the optimization of hydrophobic and hydrophilic substituent interactions of 2,4-diamino-5-(substituted-benzyl)pyrimidines with dihydrofolate reductase. *J. med. Chem.* **34**, 46–54.

Simpson, P.F. 1990 *Artificial neural systems*. Pergammon Press

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

Sternberg, M.J.E. & Thornton, J.M. 1977*b* On the conformation of proteins: an analysis of beta-pleated sheets. *J. molec. Biol.* **110**, 285–296.

Weiss, S.M. & Kulikowski, C.A. 1991 *Computer systems that learn*. San Mateo: Morgan Kaufmann.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES