Applying Inductive Logic Programming to Predicting Gene Function

Ross D. King

■ One of the fastest advancing areas of modern science is functional genomics. This science seeks to understand how the complete complement of molecular components of living organisms (nucleic acid, protein, small molecules, and so on) interact together to form living organisms. Functional genomics is of interest to AI because the relationship between machines and living organisms is central to AI and because the field is an instructive and fun domain to apply and sharpen AI tools and ideas, requiring complex knowledge representation, reasoning, learning, and so on. This article describes two machine learning (inductive logic programming [ILP])–based approaches to the bioinformatic problem of predicting protein function from amino acid sequence. The first approach is based on using ILP as a way of bootstrapping from conventional sequence-based homology methods. The second approach used protein-functional ontologies to provide function classes and a hybrid ILP method to predict function directly from sequence. Both ILP approaches were successful in producing accurate prediction rules that could biologically be interpreted. The work was also of interest to machine learning research because it highlighted the flexibility of ILP systems in dealing with heterogeneous data, the importance of problems where classes are related hierarchically, and problems where examples have more than one functional class.

We live in interesting scientific times. For the first time in history, we have access to the complete genomes of living organisms. These genomes provide the complete specification of the parts and programs to create living organisms. This knowledge is revolutionizing biology.

Has this anything to do with AI? I believe yes. The relationship between machines and living organisms is central to AI and was a core interest of the founding fathers (Alan Turing, John von Neuman, and Norbert Wiener). Modern biology is also an instructive and fun domain to apply and sharpen our tools and ideas; it requires complex knowledge representation, reasoning, learning, and so on. Because the concrete generally precedes the abstract, real-world applications catalyze new research areas.

The sequencing of a genome is only a first step to fully understanding how a living organisms works. In computer science terms, knowledge of a genome is only equivalent to obtaining a binary dump of a program, with the complication that the function and language of the program are unknown and the code undocumented and badly written. The key current scientific challenge in biology is to unravel and understand this code. AI tools will be important in this task.

Functional Genomics

Perhaps the most important discovery from the sequenced genomes is that the functions of only approximately 40 to 70 percent of the predicted genes are typically known with any confidence. For example, in bakers' yeast *(S. cerevisiae),* one of the most intensely studied of all organisms, of the approximately 6000 predicted protein-encoding genes, the function of only approximately 70 percent can be assigned with any confidence. The new science of functional genomics (Bussey 1997; Hieter and Boguski 1997) is dedicated to determining the function of genes of unassigned function and to further detailing the function of genes with purported function. Most functional genomics is concerned with the development of new experimental techniques for elucidating gene function (Blackstock and Weir 1999; DeRisi, Iyer, and Brown 1997; Oliver et al. 1998). Bioinformatics is playing an essential role in analyzing data and integrating functional genomics data (Bork et al. 1998; Brent 1999; Kell and King 2000).

A key bioinformatic problem in functional genomics is the prediction of protein function from sequence. Such predictions both provide important initial information about newly sequenced genomes and aid "wet" experimental determination of function. Such predictions are usually done by utilizing sequence similarity methods to find an evolutionary related (homologous) protein in the database that has a known function (Altschul et al. 1997; Pearson and Lipman 1988). The function of the new sequence is then inferred to be the same as the homologous protein because it is assumed to have been conserved over evolution. This inference is a kind of nearest-neighbor type in sequence space. Unfortunately, with this approach, only around 50 percent of possible homologies are identified (Park et al. 1998), and little biological insight is obtained.

Inductive Logic Programming

Machine learning and data-mining methods that utilize first-order predicate logic (FOPL) to represent background knowledge and theories are generally described as coming from the field of inductive logic programming (ILP) (Lavrac and Dzeroski 1994; Muggleton 1992, 1990) or relational data mining (RDM) (Dzeroski and Lavrac 2001), depending on the research emphasis. For simplicity in this article, I refer to all such algorithms as coming from ILP. In ILP, background knowledge (B), examples (E), and hypotheses (H) are represented as logic programs. It is first assumed that there is a requirement for an inductive hypothesis; that is, $B \not\models E$ (prior necessity). The core ILP problem is to find hypotheses H such that $B \wedge H \models E$. Matters are complicated slightly by the fact that evidence in ILP is usually divided into two types: (1) E^+ data that are consistent with the conjoined hypothesis and background information and (2) E^- data that are contrary to the conjoined hypothesis and background information. Therefore, H is required to meet the following criterion: $B \wedge H \models E^+$ and $B \models H \not\models E$. It is typical that H is also restricted to necessarily meet other requirements such as being nontrivial (not, for example, E, or $B \Rightarrow E$) and parsimonious. In practical applications, there is the additional problem that there is usually noise present (that is, some E^+ are false, and some E^- are true), which means that the previous conditions need to be relaxed such that the best hypothesis maximizes its coverage of E^+ and minimizes its coverage of E^-. These two values give a cost ratio that represents the divergence from the ideal. The problem of learning H is typically designed as a search problem through the space of models meeting the previous criteria (Mitchell 1982).

ILP has shown its value in many scientific problems such as drug design and toxicology (for example, Dzeroski et al. [1999]; Finn et al. [1998]; King, Srinivasan, and Dehaspe [2001]; King et al. [1996, 1992]) and molecular biology (Badea 2003; Donescu et al. 2002; King et al. 1994; Muggleton, King, and Sternberg 1992; Sternberg et al. 1994; Turcotte et al. 2001). ILP has been particularly well suited to problems dealing with molecular structure. In such problems, ILP has often found solutions not accessible to standard statistical, neural network, propositional machine learning, or genetic algorithms (King et al. 1996). The theories produced by ILP have also been generally more comprehensible than those using propositional methods because they are more compact and closer to natural language (King et al. 1992; Turcotte, Muggleton, and Sternberg 2001).

The Suitability of Using Inductive Logic Programming for the Prediction of Protein Function

If a problem can satisfactorily be represented and solved using propositional methods, then there is no need to apply ILP techniques. Propositional methods are generally better developed and more computationally and statistically efficient. ILP methods do not necessarily default to efficient propositional learners when given wholly propositional data. Use of ILP therefore needs to be justified. My rationale is based on the following features of the problem and the required solution:

Relational descriptors: Functional genomics naturally involves many relationships in the data: *phylogenic hierarchies* (the tree of life), *homologies* (genes sharing a common ancestor), directed graphs relating functions, and so on. Traditional propositional methods (statistical, neural network, machine learning, genetic algorithms, and so on) cannot efficiently represent these relations in inductive inference.

Data heterogeneity: The relevant data come from many different sources and are necessarily stored in multiple tables of relational databases. To use a conventional data-mining algorithm, the tables would have to be joined to form a single prohibitively large table for analysis, which is impractical. In addition, ILP/RDB allows the direct analysis of the multiple-table formatted data.

Comprehensible results: It is important that the prediction rules are understandable. Biologists generally require that the rules are understandable so that they can suggest new biological ideas and have confidence in them. In some bioinformatic applications, this requirement is not necessary, for example, in predicting protein secondary structure. However, given a choice, comprehensible results are always preferred.

At Aberystwyth, researchers have developed two approaches to applying ILP to predicting protein function. The first is called *homology induction* and is based on utilizing machine learning to improve on conventional sequence- based homology methods (Karwath and King 2002, 2001). The second method uses a hybrid ILP–propositional machine learning method to predict protein functional class directly from sequence (King, Srinivasan, and Dehaspe 2001; King et al. 2001a, 2001b, 2000).

Predicting Gene Homology

The identification of evolutionary-related (homologous) proteins is a key problem in computational molecular biology. Knowledge of a homologous relationship between two proteins, one of known function and the other of unknown function, allows the probabilistic inference that the protein with unknown function has the same function as that of the known one (because evolution generally conserves function). Such inferences are the basis of most of our knowledge about sequenced genomes. Protein homology is typically inferred by using computer programs to measure the similarity of two or more proteins. This inference is generally done by comparing the two amino acid strings of the proteins and measuring the characterwise similarity between them. Such programs probably consume more processing time than all other bioinformatic programs put together. These methods perform well for closely related homologous sequences. However, the results for more distantly related proteins are less reliable (Park et al. 1998), detecting only about 50 percent of all possible homologies, given an acceptable false-positive rate.

In learning problems, all relevant information should be used. The idea behind homology induction is to exploit additional sequence information to bootstrap on the performance of standard sequence homology methods.

Methodology

Homology induction uses background knowledge, together with the protein's amino acid sequence, to induce homology. The idea is to collect as much information as possible for a protein and then infer homology using discriminatory ILP. The homology induction approach is based on the following steps:

First is the collection of possible homologous proteins using an existing method of sequence similarity search (SSS). Aberystwyth researchers use PSI-BLAST (Altschul et al. 1997) which is essentially an iterative nearest-neighbor method (in sequence space). The result of a PSI-BLAST search is a list of possible homologous proteins sorted by probability. Proteins where the probability of homology is ambiguous are termed to be in the "twilight zone."

Second is the accumulation of all available information for these proteins. We developed a large multitable database of DATALOG (Ullman 1988) facts to describe the proteins from a wide variety of bioinformatic sources. This information was selected for relevance to the detection of homology. For each protein, we collected bioinformatic database keywords, the organism's classification (family tree), bioinformatic database references (PROSITE, HSSP, EMBL, PIR—excluding SCOP [structural classification of protein] classifications), predicted secondary structure (Ouali and King 2000), amino acid distribution for singlets and pairs of residues, and so on.

Third is the induction of rules. We used the ILP ALEPH algorithm to learn rules that were true for proteins of very high probability of being homologous (based on sequence similarity)[1] and false for proteins with close to zero probability of being homologous (also based on sequence similarity).

Fourth is the application of the rules. We applied the rules to the set of twilight zone proteins to predict whether they were homologous.

To assess the accuracy of homology induction, it was necessary to have a "gold standard" set of known homologies. We used the systematic approach of Park et al. (1997), which used a subset of the SCOP database (Murzin et al. 1995). The SCOP database is a classification database of proteins of known structure; most also have known function. Over evolutionary time, protein structure changes more slowly than sequence; therefore, structure can be used to identify more remote homologies than sequence. At the family level of SCOP, the structures are so similar that homology is inferred, which is not to be confused with protein *fold recognition*, where there is no necessary expectation of homology.

A perfect prediction method would be able to detect all homologous relationships in SCOP. However, in practice, unrelated nonhomologous proteins are predicted (errors of commission), and evolutionary related proteins are missed (errors of omission). The cost of these different types of errors depends on the biological problem. Therefore, we used receiver-operating characteristic (ROC) curves to compare our predictions (Bradley 1995).

Results

Homology induction induced rules for 1,015 proteins of known structure (PDB40D). The original PSI-BLAST results were used for the sequences where no rules could be induced. In total, homology induction produced 1,851 rules. The most commonly used predicate of the single predicate rules was db_ref, used by 651 rules. These rules consisted mainly of references to the bioinformatic database PROSITE.[2] This result was expected because this database contains patterns designed to cluster homologous families of proteins together. In the PDB40D database, there are 8,022 true homology relationships and 2,046,900 false ones. The accuracy for PSI-BLAST was 99.69 percent and for homology induction 99.70 percent. Although the accuracy of homology induction is marginally higher than PSI-BLAST alone, it is not clear at first sight if it is significantly higher. To test significance, researchers therefore performed a two-sample χ^2 test to compare the actual frequency of a prediction with the estimated frequency of the prediction. The critical value of χ^2 for 1 degree of freedom and 99.995-percent confidence is 7.879, which indicates that homology induction was significantly better than PSI-BLAST alone.

For comparison over all linear costs, I performed a ROC analysis. We compared the area under the ROC (AUROC) for homology induction and PSI-BLAST. The AUROC value for homology induction was 0.65, and the AUROC for PSI-BLAST 0.61. The ROC curve for PSI-BLAST, along with the ROC curves for the standard homology induction (HI^{all}) and a version of homology induction based on the subset of descriptors directly calculable from only sequence information (HI^{seq}) is shown in figure 1. The dominating curve is that of HI^{all}, to the left of the other two curves. The ROC curve of HI^{seq} does not entirely dominate the PSI-BLAST ROC curve, and for large sections of the false positive axis, the two curves have a similar true positive rate. However, for the false positive rate interval of 0.38 to 0.5, the ROC curve produced by HI^{seq} does clearly dominate that of PSI-BLAST.

To illustrate the biological utility of the homology induction rules, I use the protein C-Phycocyanin (1CPC). Figure 2 shows the HI^{all} and HI^{seq} rules learned for C-Phycocyanin both in their original Prolog form and in English translation. Phycocyanins are light harvesting proteins. Applying PSI-BLAST to the data produced three proteins in the twilight zone: (1) allophycocyanin alpha-b chain *(Anabena),* (2) erythroid transcription factor (gata-1 *Mus musculus*), and (3) oryzain gamma chain precursor *(Oryza sativa)*. The rules in the HI^{all} and HI^{seq} rule sets correctly identified the allophycocyanin as homologous to 1CPC. There is convincing experimental evidence for this homology and note that PSI-BLAST does not use protein names! No rule in any rule set identified the other two twilight zone sequences as homologous, which would appear to be correct (no structures exist to be certain). Further evidence for the power of the homology induction rules is that the homology induction analysis was applied to version 37.0 of the bioinformatic database SWISS-PROT,[3] and each of the 13 positive examples not covered by this rule have had the keyword *phycobilisome* added to their annotation since version 38.0 of SWISS-PROT. It is particularly intriguing that the most characteristic feature of the amino acid–type rules is low-histidine and -trypotophan content and that both amino acids have nitrocyclic aromatic rings, which can be explained chemically. Phycocyanins have covalently linked bilin prosthetic groups that consist of linked nitrocyclic aromatic rings. Aberystwyth researchers hypothesize that evolution has selected for low-histidine and -trypotophan content in phycocyanins to reduce electron-transport interference. The requirement for a high number of leucine-arginine pairs is also structurally significant because these arginines form salt bridges with the prosthetic groups. The structural rule *s*2 is also consistent with the known structure of phycocyanins, which are well known to have an all α-helix globinlike fold.

Predicting Protein Functional Class

Perhaps the most important recent advance in bioinformatics has been the development of good ontologies to describe protein function, for example, GO and RILEY.[4,5] These ontologies take the form of hierarchies or directed acyclic graphs. Figure 3 illustrates part of the Riley hierarchy for the bacteria *E. coli*, one of the best-established functional ontologies. The creation of such ontologies opened up the possibility of directly predicting protein func-

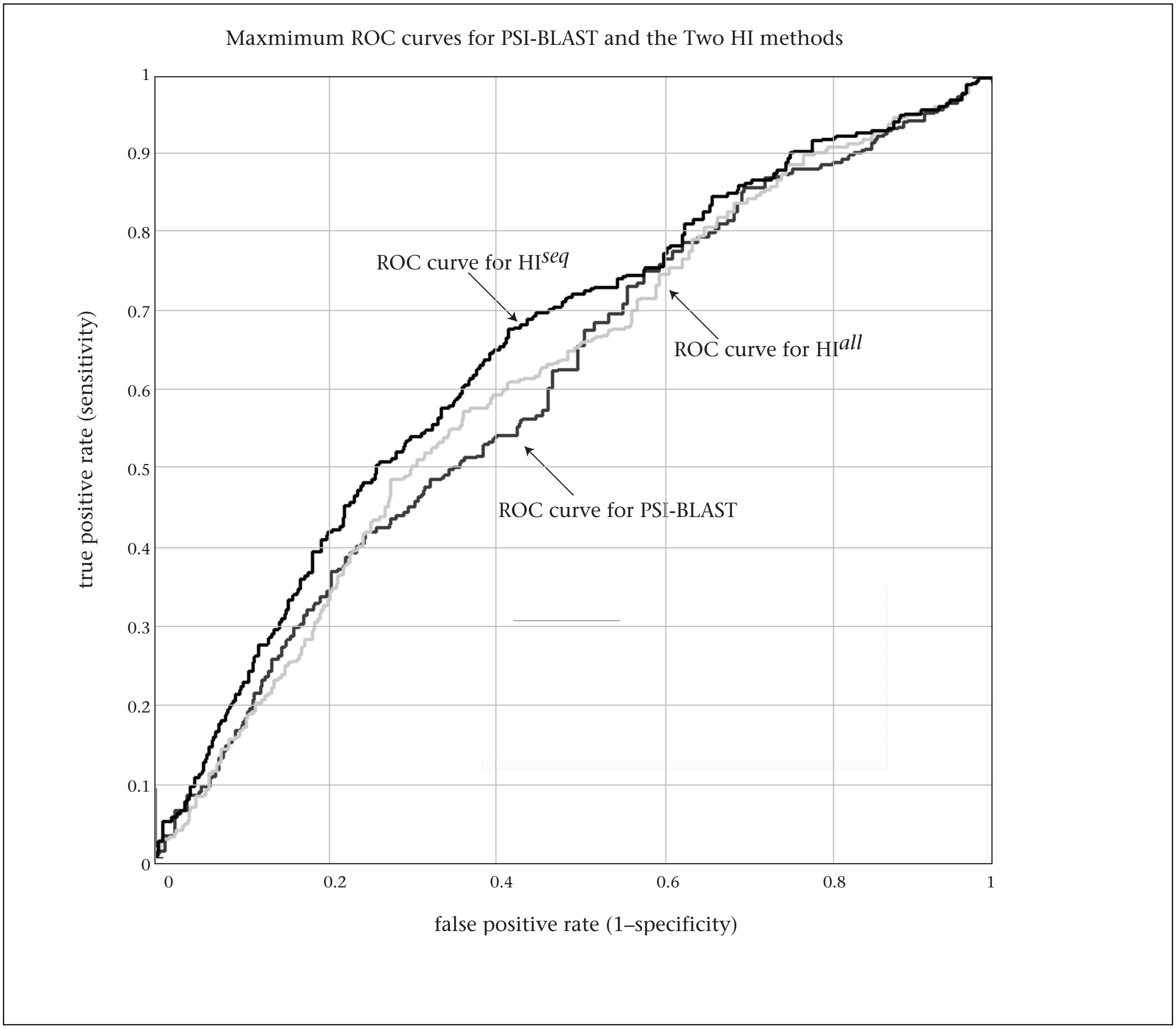

Figure 1. The Three Receiver-Operating Characteristic (ROC) Curves Produced by PSI-BLAST, HI^{ALL} and HI^{seq}, for Predictions in the Twilight Zone.

Although the ROC curve for PSI-BLAST results from applying ROC analysis directly to the results produced, the ROC curves for both homology induction methods are maximized using a cross-validated value for resorting. The ROC curve for HI^{all} dominates over the other two curves at all times, but the curves for PSI-BLAST and HI^{seq} oscillate around each other. HI^{seq} dominates the PSI-BLAST curve between ~0.38 and ~0.5.

tional class from sequence. Abstractly, what is required is a discrimination function that maps sequence to biological functional class. The existing sequence homology recognition methods (see earlier discussion) can be viewed as examples of such functions: Methods based on direct sequence similarity can be considered as nearest-neighbor–type functions (in sequence space), and the more complicated homology recognition methods based on motifs and profiles resemble case-based learning methods.

Methodology

We selected the *E. coli* genome to test the idea of using machine learning to learn predictive mappings between protein sequence and function. *E. coli* is probably the best characterized extant genome and is the "model" bacteria. It has an estimated 4,289 identified proteins (Blattner et al. 1997). Of these proteins, approximately 30 percent had unknown function. To predict functional class, we collected similar data to that used in homology induc-

PDB 1CPC C-Phycocyanin

HI^{all}

Prolog

```
        homologous(A) :-
                        desc(A,chain),
                        amino_acid_ratio_rule(A,h,1).
        homologous(A) :-
                        keyword(A,phycobilisome).
```

English

```
        A protein is homologous if
a1              it has the word “chain” in its SWISS-PROT description line and
                it has a level 1 histidine content in the residue chain and
a2      or      it has the word “phycobilisome” as a SWISS-PROT  keyword.
```

HI^{seq}

Prolog

```
        homologous(A) :-
                        amino_acid_ratio_rule(A,w,1),
                        amino_acid_ratio_rule(A,h,1),
                        amino_acid_pair_ratio_rule(A,l,r,10).
        homologous(A):-
                        mol_wt_rule(A,3),
                        sec_struc_distribution_rule(A,a,10).
```

English

```
        A protein is homologous if
s1              it has a level 1 tryptophan content and
                it has a level 1 histidine content and
                it has a level 10 leucine-arginine pair content.
s2      or
                it has a level 3 molecular weight and
                it has a level 10 predicted α-helix content.
```

Figure 2. The Homology Induction Rules Learned to Identify 1CPC (C-Phycocyanin) Are Illustrated First in Their Original Prolog Form and Then in English Translation.

Two sets of rules are shown, those using HI^{all} and those learned from HI^{seq}.
All numbers were discretized into 10 levels for ease of symbolic induction (1 low to 10 high).

tion (see earlier discussion). We formed a DATALOG database containing all the data we could find on the protein sequences. The most commonly used technique to gain information about a sequence is to run a sequence similarity search, which was used as the starting point in forming descriptions (we used PSI-BLAST). For each protein in the genome, we formed a description based on the frequency of singlets and pairs of residues in the protein, the phylogeny (family tree) of the organism from which each homologous protein was obtained from SWISS-PROT; SWISS-PROT protein keywords from homologous proteins, the length and molecular weight of the protein, and its predicted secondary structure using PROF (Ouali and King 2000). In total, 10,097,865 DATALOG facts were generated for the *E. coli* genome.

To analyze this database, we used a hybrid combination of ILP and propositional tree learning (figure 4). The ILP data-mining program WARMR (Dehaspe and Toivonen 1999) was first used to identify frequent patterns (conjunctive queries) in the databases. WARMR is a

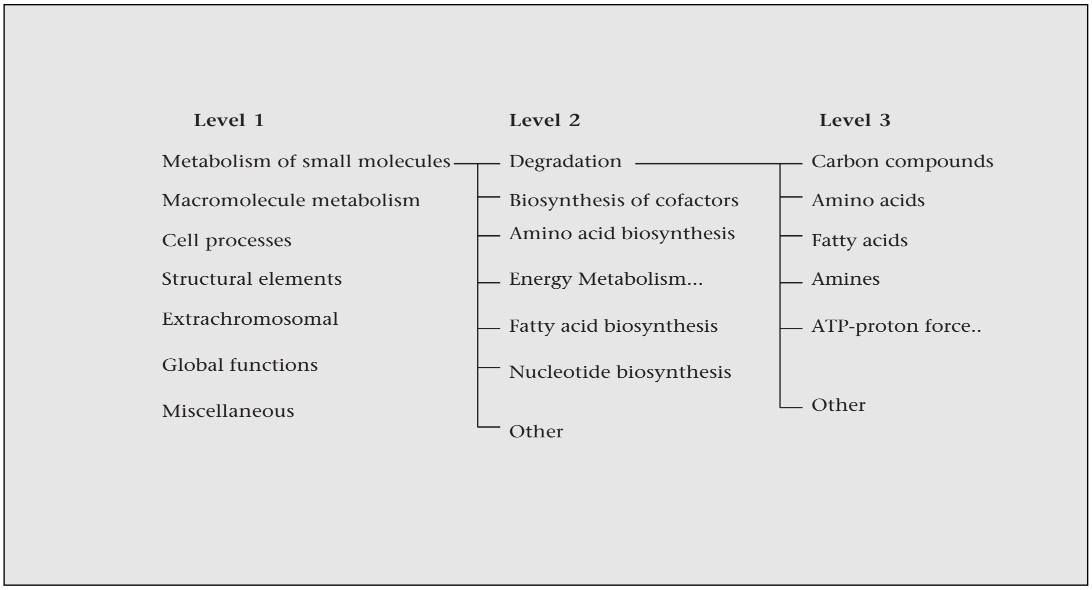

Figure 3. An Example Subset of the Riley Group Protein Functional Ontology in E. coli.

	Level 1	Level 2	Level 3
No. of rules found	13	13	13
No. predicting more than one homology class	9	10	3
No. predicting a new homology class	9	5	3
Average test accuracy	75%	69%	61%
Default test accuracy	40%	21%	6%
New functions assigned	353 (16%)	267 (12%)	135 (6%)

Table 1. Learning Results for E. coli.

The number of rules found are those selected on the validation set. A rule predicts more than one homology class if there is more than one sequence similarity cluster in the correct test predictions. A rule predicts a new homology class if there is a sequence similarity cluster in the test predictions that has no members in the training data. Average test accuracy is the accuracy of the predictions on the test proteins of assigned function (if conflicts occurred; the prediction with the highest a priori probability was chosen). Default test accuracy is the accuracy that could be achieved by always selecting the most populous class. *New functions assigned* is the number of proteins of unassigned function predicted. The test accuracy estimates might be too pessimistic because proteins can have more than one functional class, but only one of these is considered correct.

general-purpose data-mining algorithm that can discover knowledge in structured data. It can learn patterns reflecting one-to-many and many-to-many relationships over several tables. No standard data-mining program can do this because they are restricted to simple associations in single tables. WARMR uses a first-order version of the efficient levelwise a priori algorithm (Agrawal and Srikant 1994), which allows it to be used on very large databases. The WARMR levelwise search algorithm is based on a breadth-first search of the pattern space. The application of WARMR can be considered as a way of identifying the most important structure in a database. In the *E. coli* database, WARMR discovered approximately 18,000 frequent queries.

These frequent patterns were converted into Boolean (indicator) attributes for propositional rule learning. An attribute has value 1 for a specific gene if the corresponding query succeeds for that gene and 0 if the query fails. The propositional machine learning algorithm C5 (Quinlan 1993) was then used to induce rules that predict function from these Boolean attributes. Good rules were selected on a validation set and the unbiased accuracy of these rules estimated on a test set. Rules were selected to balance accuracy with unidentified gene coverage. The prediction rules were then ap-

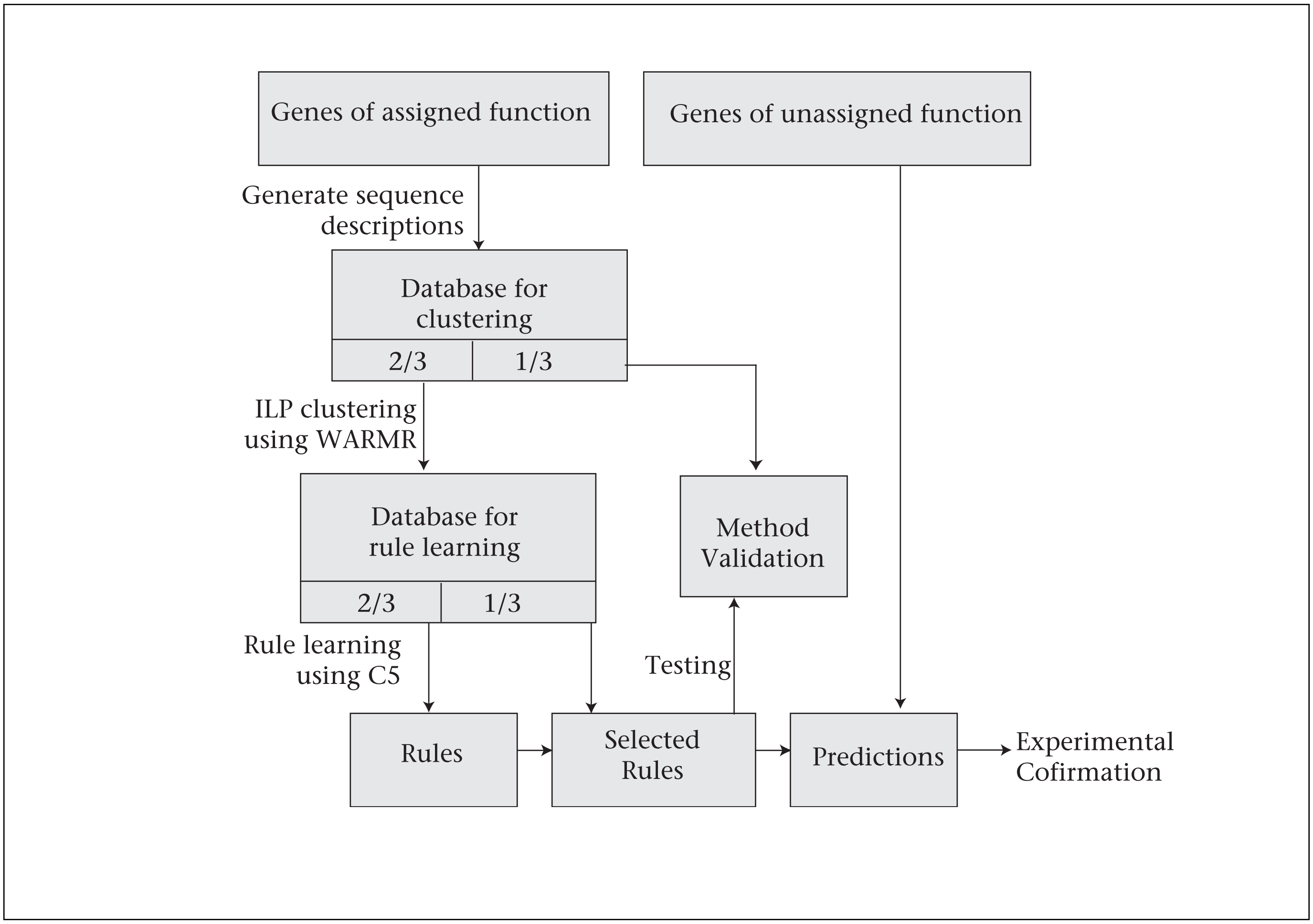
Genes of assigned function
Genes of unassigned function
Generate sequence descriptions
Database for clustering
2/3
1/3
ILP clustering using WARMR
Database for rule learning
2/3
1/3
Method Validation
Rule learning using C5
Testing
Rules
Selected Rules
Predictions
Experimental Cofirmation

Figure 4. Flowchart of the Data-Mining Program Methodology.

This inductive logic programming (ILP)/propositional hybrid approach has proved successful in the past on other scientific discovery tasks. It is powerful because the clustering improves the representation for learning (using the expressive power of ILP), and the discrimination step efficiently exploits the prelabeled examples. Good rules were selected on a validation set and the unbiased accuracy of these rules estimated on the test set. The unbiased accuracy of these rules was estimated on the 1/3 test set. The selection criteria for good rules was that on the validation data, they covered at least two correct examples, had an accuracy of at least 50 percent, and had an estimated deviation of ≥ 1.64.

plied to genes that have not been assigned a function to predict their functions. **Note:** We did not aim for a general model of the relationship between sequence and function; we were satisfied with finding good rules to cover part of the space.

Results

In 2000 to 2001, we published this data-mining prediction approach to predicting protein functional class from sequence (King, Srinivasan, and Dehaspe 2001; King et al. 2001a, 2000b, 2000). By using a held-out test set of proteins with annotated function, data-mining prediction had an estimated accuracy of 50 to 90% percent (depending on the position of class in the function ontology). A summary of these results is given in table 1. A key finding was that the method could learn predictive rules that were more general than is possible using homology-based methods.

Using the same prediction rules, we also predicted the functional class of 1,309 proteins of then-unknown function. These predictions were made publicly available at the gene-predictions web site.[6] Statistical theory, and the design of our machine methodology, gave us confidence in these predictions. However, doubts remained: It seemed a priori unlikely that protein function could be predicted from sequence (predicting protein structure from sequence has proved intractable, and predicting function from structure is an unsolved problem), and it was possible that the proteins of unknown function came from a significantly different distribution from those of known function—which would invalidate a key statistical–machine learning assumption.

If The ORF is not predicted to have a β-strand length ≤ 3 **and**
a homologous protein from class *Chytridiomycetes* was found
Then its functional class is “Cell processes, Transport/binding proteins”

Figure 5. A Level-2 Rule.

This rule is based on predicted structural and phylogenic features. In the original test set, this rule was 12/13 (86 percent) correct. The default accuracy for this class is 21 percent. Twenty-four genes of unknown function were predicted by the rule. Of these 24 predictions, 2 have been confirmed by experiment, 7 have been (independently) annotated to have the predicted function, and 1 has been annotated to have a nonpredicted function. The rule also has a possible biological explanation. We hypothesize that cytochrome c oxidase in *Chytridiomycetes* is a “molecular living fossil” that has retained features of an ancestral protein that radiated into a wide variety of transport proteins, which has allowed the protein to be used to identify very remote homologous that would otherwise be missed.

In the period since these predictions, biological knowledge has advanced greatly. Some proteins in *E. coli* have had their function determined by wet biology. Function determination has also occurred in other organisms, allowing better homology-based function predictions. Equally important, many more protein sequences have been determined, allowing sequence-similarity methods to predict function with greater accuracy. This new biological knowledge allows us to test our predictions directly. We used two ways to test the predictions.

First, we compared our predictions to the updated (20.02.02) Monica Riley genome group annotation, which has the advantage of testing a large number of predictions.

Second, we examined the scientific literature for the direct experimental derivation of protein functions for our predictions. This test has the advantage of directly testing the predictions experimentally.

To test for the probability of our predictions occurring by chance, Aberystwyth researchers used a binomial test, with the probability of success being the probability of the most populous class. This test has the advantage of being simple to calculate, makes few assumptions, and is guaranteed to give an overestimate.

The results for the new Riley group annotation were statistically highly significant (< 1e-15), with prediction accuracies of approximately 90 percent for the cases where more than one rule agreed on a prediction. It should also be stressed again that these accuracies are likely to be underestimates because they are based on the assumption that the Riley annotation is complete and correct.

The results for the function predictions that have either been confirmed, or not, by wet biological experiments were also highly significant, although at lower accuracy than for the annotations (probably because of bias in the sample of functions confirmed). See table 2 for details of the results for level 3 of the function hierarchy. An example prediction rule is shown in figure 5. It was gratifying that the rules illustrated in previous publications were found to have accuracies consistent with prediction on this blind trial.

Discussion

In this article, I described two applications of the first-order machine learning methodology, inductive logic programming, to the problem of predicting protein function from sequence.

I consider biology first. Are the results of any practical use? On this question, I believe the jury is still out. The results of homology induction (although a statistically significant improvement on PSI-BLAST) are perhaps too small a step-change to make biologists use the system en masse. However, the functional class results are probably more significant. These I believe constitute a step-change in protein function prediction methodology. Although initially, many biologists were skeptical, this reaction seems to be slowly changing, and interest is growing in the approach. The new evidence of the results of the blind-test predictions (table 2) should help the acceptance of this methodology. One important limitation of the data-mining–prediction approach is that although the rules are presented in a symbolic form, their meaning is often obscure. It is certainly possible to find biological justification for some of the rules (see, for example, figure 5), but in many cases, the biological meaning is obscure, even when the rules are empirically successful. Much more work is needed on the design of learning systems that produce semantically comprehensible results.

A number of interesting machine learning issues were brought into focus by the applications:

ORF	Rule	Predicted Class	Confirmed Function	Result
b0533	104	Surface structures	Fimbrial assembly protein (ADHESION)	C
b0570	56	Global regulatory functions	Multimodular CusS: sensory kinase in regulatory system	C
b0613	83	Conversions of intermediate metabolism	2-(5''-triphosphoribosyl)-3'-dephosphocoenzyme-A synthase	C
b0619	56a	Global regulatory functions	Sensory histidine kinase in two-component reg. system	C
b0619	56b	Global regulatory functions	Sensory histidine kinase in two-component reg. system	C
b1981	63	MFS family	ShiA: MFS family	C
b1981	66	MFS family	ShiA: MFS family	C
b1981	108	MFS family	MFS family, shikimate	C
b2219	56	Global regulatory functions	Sensory protein kinase in two-component reg. system	C
b2972	62	Chemotaxis and mobility	Bifunctional prepilin peptidase	C
b0053	39	Degradation of DNA	Peptidyl-prolyl cis-trans isomerase	W
b0162	107	Transposon-related functions	Regulator of D-galactarate, D-glucarate and D-glycerate metabolism	W
b0441	39	Degradation of DNA	Peptidyl-prolyl cis-trans isomerase	W
b0505	108	Transposon-related functions	Ureidoglycolate hydrolase	W
b0508	94	Ribosomal proteins	Hydroxpyruvate isomerase	W
b0662	56	Global regulatory functions	Oxygenase involved in ubiquinone biosynthesis	W
b0789	108	Transposon-related functions	Cardiolipin synthase activity	W
b1199	147	Transposon-related functions	DHA kinase domain	W
b2052	106	Transposon-related functions	Bifunctional GDP-fucose synthetase	W
b3338	108	MFS family	Periplasmic endochitinase	W
b3419	142	Surface structures	RNA 3'-terminal phosphate cyclase	W
b3836	107	Transposon-related functions	Component of translocase	W
b3838	106	Transposon-related functions	Essential component of translocase	W
b2392	62	Chemotaxis and mobility	High-affinity manganese transporter	W
b2392	2	ABC superfamily (membrane)	High-affinity manganese transporter	NM
b2392	97	STP family	High-affinity manganese transporter	NM

Table 2. Predictions of Classes in Level 3 of the E. coli Gene Ontology That Now Have Wet Biological Evidence.

ORF is the Blattner identifier for the protein. The predictions are ordered by result and ID. The rule numbers are identifiers for the specific rule predicting the gene. *C* = Correct, *W* = Wrong, *NM* = Near Miss. There are 10 correct predictions and 14 wrong ones. The probability of obtaining this accuracy on newly determined functions occurring by chance is estimated at less than 4.8e-10.

Was the use of ILP required? In many machine learning applications, perhaps most, it is not necessary to use ILP/relational data mining because propositional methods are sufficient because there has been orders-of-magnitude more work done on propositional methods, and ILP methods do not necessarily act as efficient propositional learners when given wholly propositional data. For example, in bioinformatics, propositional methods would empirically seem sufficient to predict protein secondary structure because neural network approaches have time after time been the most successful in blind trials.[7] However, in the prediction of protein function, it is very hard to see how the crucial relational aspects of the problem could be encoded efficiently.

The functional classes for proteins exist in hierarchies or directed acyclic graphs, which means that the classes are not independent of each other. Problems with this characteristic are relatively common in the real world (for example, in text classification) but have been little considered by the statistical or machine learning community (Clare and King 2001; Kohler and Sahami 1997).

It is possible for proteins to have more than one function, that is, to have more than one class value. Such problems are also common in the real world and little studied (for example, Clare and King [2001]; Schapire and Singer [2000]). Of course, it is always possible to create disjoint classes, but this can distort the problem and create large numbers of artificial classes.

In conclusion, the application of AI to deciphering genomic information is only just beginning. Enormous challenges exist in data integration, the analysis of data from microarrays, proteomics, and so on. The dream for the future is to be able to develop models of cells, development, tissues, and even whole organisms. AI has the potential to contribute substantially to this enterprise. In turn, AI will greatly gain in the process.

Acknowledgments

I would like to thank Amanda Clare, Luc Dehaspe, and Andreas Karwath who worked with me on homology induction and data- mining prediction as well as Douglas Kell, Walter Luyten, and Mike Young for helpful comments on the work. The work was partly funded by Engineering and Physical Science Research Council grant GR/L62849 and Medical Research Council grant G78/6609.

Notes

1. web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph.
2. ca.expasy.org/prosite.
3. www.ebi.ac.uk/swissprot.
4. www.geneontology.org.
5. genprotec.mbl.edu/riley.
6. www.aber.ac.uk/compsci/Research/bio/ProteinFunction.
7. CASP: predictioncenter.llnl.gov.

References

Agrawal, R., and Srikant, R. 1994. Fast Algorithms for Mining Association Rules in Large Databases. Paper presented at the Twentieth International Conference on Very Large Databases (VLDB), 12–15 September, Santiago, Chile.

Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acid Research 25(17): 3389–3402.

Badea, L. 2003. Functional Discrimination of Gene Expression Patterns in Terms of the Gene Ontology. Paper presented at the Pacific Symposium on Biocomputing, 3–7 January, Lihue, Hawaii.

Blackstock, W. P., and Weir, M. P. 1999. Proteomics: Quantitative and Physical Mapping of Cellular Proteins. *TIBTECH* 17(1): 121–127.

Blattner, F. R.; Plunkett, G., 3d.; Block, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; and Shao, Y. 1997. The Complete Genome Sequence of *Escherichia coli* K–12. *Science* 277(5331): 1432–1434.

Bork, P.; Dandekar, T.; Diaz-Lazcoz, Y.; Eisenhaber, F.; Huynen, M.; and Yuan, Y. P. 1998. Predicting Function: From Genes to Genomes and Back. *Journal of Molecular Biology* 283(4): 707–725.

Bradley, A. P. 1995. The Use of Area under ROC Curve in the Evaluation of Learning Algorithms. *Pattern Recognition* 30(6): 1145–1159.

Brent, R. 1999. Functional Genomics: Learning to Think about Gene Expression Data. *Current Biology* 9(9): R338–R341.

Bussey, H. 1997. 1997 Ushers in an Era of Yeast Functional Genomics. *Yeast* 13(16): 1501–1503.

Clare, A. J., and King, R. D. 2001. Knowledge Discovery in Multilabel Phenotype Data. In *Proceedings of the Fifth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01),* eds. L. De Raedt and A. Siebes, 42–53. Lecture Notes in Artificial Intelligence 2168. Heidelberg, Germany: Springer-Verlag.

Dehaspe, L., and Toivonen, H. 1999. Discovery of Frequent DATALOG Patterns. *Data Mining and Knowledge Discovery* 3(1): 7–16.

DeRisi, J. L.; Iyer, V. R.; and Brown, P. O. 1997. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* 278(5338): 680–686.

Donescu, A.; Waissman, J.; Richard, G.; and Roux, G. 2002. Characterization of Bio-Chemical Signals by Inductive Logic Programming. *Knowledge-Based Systems* 15:129–137.

Dzeroski, S., and Lavrac, N. 2001. *Relational Data Mining.* Berlin: Springer-Verlag.

Dzeroski, S.; Blockeel, H.; Kompare, B.; Kramer, S.; Pfahringer, B.; and Van Laer, W. 1999. Experiments in Predicting Biodegradability. In *Proceedings of the Ninth International Workshop on Inductive Logic Programming,* 80–91. Lecture Notes in Artificial Intelligence 1634. New York: Springer-Verlag.

Finn, P.; Muggleton, S.; Page, D.; and Srinivasan, A. 1998. *A. Pharmacophore* Discovery Using the Inductive Logic Programming System PROGOL. *Machine Learning* 30(2): 241–271.

Hieter, P., and Boguski, N. 1997. Functional Genomics: It's All How You Read It. *Science* 278(5338): 601–602.

Karwath, A., and King, R. D. 2002. Homology Induction: The Use of Machine Learning to Improve Sequence Similarity Searches. *BMC Bioinformatics* 3(1): 11.

Karwath, A., and King, R. D. 2001. An Automated ILP Server in the Field of Bioinformatics. In *Proceedings of the Eleventh International Conference on Inductive Logic Programming (ILP'01),* eds. C. Rouveirol and M. Sebag, 91–103. Lecture Notes in Artificial Intelligence 2157. Heidelberg: Springer-Verlag.

Kell, D., and King, R. D. 2000. On the Optimization of Classes for the Assignment of Unidentified Reading Frames in Functional Genomics Programmes: The Need for Machine Learning. *Trends in Biotechnology* 18(3): 93–98.

King, R. D.; Clark, D. A.; Shirazi, J.; and Sternberg, M. J. E. 1994. On the Use of Machine Learning to Identify Topological Rules in the Packing of Beta-Strands. *Protein Engineering* 7(11): 1295–1303.

King, R. D.; Srinivasan, A.; and Dehaspe, L. 2001. WARMR: A Data-Mining Tool for Chemical Data. *Journal of Computer-Aided Molecular Design* 15(2): 173–181.

King, R. D.; Karwath, A.; Clare, A.; and Dehapse, L. 2001. The Utility of Different Representations of Protein Sequence for Predicting Functional Class. *Bioinformatics* 17(5): 445–454.

King, R. D.; Karwath, A.; Clare, A.; and Dehapse, L. 2000a. Accurate Prediction of Protein Class in the *M. tuberculosis* and *E. coli* Genomes Using Data Mining. *Yeast* (Comparative and Functional Genomics) 17(4): 283–293.

King, R. D.; Karwath, A.; Clare, A.; and Dehapse, L. 2000b. Genome-Scale Prediction of Protein Functional Class from Sequence Using Data Mining. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, eds. R. Ramakrishnan, S. Stolfo, R. Bayardo, and I. Parsa, 384–389. New York: Association for Computing Machinery.

King, R. D.; Muggleton, S. H.; Srinivasan, A.; and Sternberg, M. J. E. 1996. Structure-Activity Relationships Derived by Machine Learning: The Use of Atoms and Their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming. *Proceedings of the National Academy of Sciences* 93(1): 438–442.

King, R. D.; Muggleton, S.; Lewis, R. A.; 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 Analogs Binding to Dihydrofolate-Reductase. *Proceedings of the National Academy of Sciences* 89(23): 11322–11326.

Kohler, D., and Sahami, M. 1997. Hierarchically Classifying Documents Using Very Few Words. Paper presented at the Fourteenth International Conference of Machine Learning, 8–12 July, Nashville, Tennessee.

Lavrac, N., and Dzeroski, S. 1994. *Inductive Logic Programming: Techniques and Applications*. Chichester, U.K.: Ellis Horwood.

Mitchell, T. M. 1982. Generalization as Search. *Artificial Intelligence* 18(2): 203–226.

Muggleton, S. H. 1992. *Inductive Logic Programming.* San Diego, Calif.: Academic.

Muggleton, S. H. 1990. Inductive Logic Programming. *New Generation Computing* 8(4): 295–318.

Muggleton, S.; King, R. D.; and Sternberg, M. J. E. 1992. Protein Secondary-Structure Prediction Using Logic-Based Machine Learning. *Protein Engineering* 5(7): 647–657.

Murzin, A. G.; Brenner, S. E.; Hubbard, T. J. P.; and Chothia, C. 1995. SCOP: A Structural Classification of Protein Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology* 247(4): 536–540.

Oliver, S. G.; Winson, M. K.; Kell, D. B.; and Baganz, F. 1998. Systematic Functional Analysis of the Yeast Genome. *Trends in Biotechnology* 16(10): 373–378.

Ouali, M., and King, R. D. 2000. Cascaded Multiple Classifiers for Secondary-Structure Prediction. *Protein Science* 9(6): 1162–1176.

Park, J.; Teichmann, S. A.; Hubbard, T.; and Chothia, C. 1997. Intermediate Sequences Increase the Detection of Homology between Sequences. *Journal of Molecular Biology* 273(1): 349–354.

Park, J.; Karplus, K.; Barrett, C.; Hughey, R.; Haussler, D.; Hubbard, T.; and Chothia, C. 1998. Sequence Comparisons Using Multiple Sequences Detect Three Times as Many Remote Homologues as Pairwise Methods. *Journal of Molecular Biology* 284(4): 1201–1210.

Pearson, W. R., and Lipman, D. J. 1988. Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Science* 85(8): 2444–2448.

Quinlan, R. 1993. *C4.5: Programs for Machine Learning*. San Francisco, Calif.: Morgan Kaufmann.

Schapire, R., and Singer, Y. 2000. BOOSTEXTER: A Boosting-Based System for Text Categorization. *Machine Learning* 39(2): 135–168.

Sternberg, M. J. E.; King, R. D.; Lewis, R. A.; and Muggleton, S. 1994. Application of Machine Learning to Structural Molecular Biology. *Philosophical Transactions of the Royal Society of London Series B—Biological Sciences* 344:365–371.

Turcotte, M.; Muggleton, S. H.; and. Sternberg, M. J. E. 2001. Automated Discovery of Structural Signatures of Protein Fold and Function. *Journal of Molecular Biology* 306(3): 591–605.

Ullman, J. D. 1988. *Principles of Databases and Knowledge-Based Systems, Volume 1.* Rockville, Md.: Computer Science Press.

Ross D. King completed his Ph.D. at the Turing Institute in Glasgow, Scotland. He then was a visiting professor at Denver University and head of the Biotechnology Department, Brainware GmbH, Berlin. He returned to the United Kingdom to work on the STATLOG project. He then moved to London to work in the biomolecular structure group of the Imperial Cancer Research Fund. In 1997, he moved to the Department of Computer Science at the University of Wales, Aberystwyth. He was awarded his chair in 2002. His e-mail address is rdk@aber.ac.uk.