Induction of enzyme classes from biological databases

Stephen Muggleton, Alireza Tamaddoni-Nezhad and Hiroaki Watanabe

Department of Computing,
Imperial College
University of London
180 Queen's Gate, London SW7 2BZ, UK
Email: {shm, atn, hw3}@doc.ic.ac.uk

Abstract. Bioinformatics is characterised by a growing diversity of large-scale databases containing information on genetics, proteins, metabolism and disease. It is widely agreed that there is an increasingly urgent need for technologies which can integrate these disparate knowledge sources. In this paper we propose that not only is machine learning a good candidate technology for such data integration, but Inductive Logic Programming, in particular, has strengths for handling the relational aspects of this task. Relations can be used to capture, in a single representation, not only biochemical reaction information but also protein and ligand structure as well as metabolic network information. Resources such as the Gene Ontology (GO) and the Enzyme Commission (EC) system both provide isa-hierarchies of enzyme functions. On the face of it GO and EC should be invaluable resources for supporting automation within Functional Genomics, which aims at predicting the function of unassigned enzymes from the genome projects. However, neither GO nor EC can be directly used for this purpose since the classes have only a natural language description. In this paper we make an initial attempt at machine learning EC classes for the purpose of enzyme function prediction in terms of biochemical reaction descriptions found in the LIGAND database. To our knowledge this is the first attempt to do so. In our experiments we learn descriptions for a small set of EC classes including Oxireductase and Phosphotransferase. Predictive accuracy are provided for all learned classes. In further work we hope to complete the learning of enzyme classes and integrate the learned models with metabolic network descriptions to support "gap-filling" in the present understanding of metabolism.

1 Introduction

Within Bioinformatics there is a growing diversity of large-scale databases containing information on gene sequences (eg. EMBL[1]), proteins (eg. Swiss-Prot[2],

[1] `http://www.ebi.ac.uk/embl/`
[2] `http://www.ebi.ac.uk/swissprot/`

SCOP[3]), metabolism (eg KEGG[4], WIT[5] and BRENDA[6]) and disease (eg. JG-SNP database[7]). It is widely agreed that there is an increasingly urgent need for technologies which integrate these disparate knowledge sources. In this paper we propose that not only is machine learning a good candidate technology for such data integration, but Inductive Logic Programming, in particular, has strengths for handling the relational aspects of this task. In the context of protein structure prediction this approach has already shown success [6]. Potentially relations can be used to capture, in a single representation, not only biochemical reaction information but also protein and ligand structure as well as metabolic network information.

Bioinformatic resources such as the Gene Ontology (GO) [3] and the Enzyme Classification (EC) [5] system both provide isa-hierarchies of enzyme functions. On the face of it GO and EC should be invaluable for supporting automation within Functional Genomics [8], which aims at predicting the function of unassigned enzymes from the genome projects. However, neither GO nor EC can be directly used for this purpose since the classes have only a natural language description. In this paper we make an initial attempt at machine learning logic programs describing EC classes for the purpose of enzyme function prediction. The function of any particular enzyme is normally described in terms of the biochemical reaction which it catalyses. The LIGAND database[8] provides an extensive set of biochemical reactions underlying KEGG (the Kyoto Encyclopedia of Genes and Genomes). In LIGAND reactions are described in equational form as follows.

$$C_{21}H_{30}N_7O_{17}P_3 + C_7H_8O_5 \rightleftharpoons C_{21}H_{29}N_7O_{17}P_3 + C_7H_{10}O_5$$

To our knowledge the experiments described in this paper are the first attempt to learn enzyme functions in terms of these underlying reaction equations. In our experiments we learn descriptions for a small set of EC classes including Oxireductase and Phosphotransferase. Predictive accuracies are provided for all learned classes. In further work we hope to complete the learning of GO classes and integrate the learned models with metabolic network descriptions to support "gap-filling" in the present understanding of metabolism. This will extend previous research described in [2] by allowing for the case in which substrates and products of a reaction are known, but there is no known enzyme for such a reaction. In this case, rather than hypothesis an arbitrary unknown enzyme we could potentially use abduction together with learned biochemical knowledge of the kind we develop in this paper to narrow down the functional class of the missing enzyme.

This paper is arranged as follows. Section 2 introduces the EC classification system. The LIGAND database of biochemical reactions is then described in

[3] http://scop.mrc-lmb.cam.ac.uk/scop/
[4] http://www.genome.ad.jp/kegg/
[5] http://wit.mcs.anl.gov/WIT2/
[6] http://www.brenda.uni-koeln.de/
[7] http://www.tmgh.metro.tokyo.jp/jg-snp/
[8] http://www.genome.ad.jp/ligand/

Section 3. In the experiments described in Section 4 we investigate the possibility of learning EC classification rules in terms of the biochemical reactions found in LIGAND. In Section 5 we discuss the representation issues with the learned rules. In Section 6 we conclude and describe directions for further research.

2 Enzyme Classification

Enzymes are proteins which catalyse biochemical reactions within organisms. The description of enzyme function provides a characterisation of biological systems which forms a bridge between the micro-level and the macro-level (from atoms, through chemical reaction networks, to diseases). The genome projects are generating ever-larger volumes of genes with unassigned function. This has led in turn to an increasingly important role for enzyme classification systems. Let us consider homology-based functional genomics as an approach to finding the functions of unknown enzymes using an enzyme classification system. Assume that we have an amino-acid sequence of an unknown enzyme A and a known enzyme B, and that B belongs to the enzyme class X. First, we use software to compute the amino-acid sequence similarity of A and B in order to determine their degree of homology. If A and B are found to be homologues then we next proceed with experiments based on the hypothesise that A has a similar function to B.

Classification of enzyme functions is difficult since many enzyme mechanisms are not fully understood and many enzymes catalyse multiple reactions. To tackle this issue, classification systems have been proposed which focus on different features of enzymes [5, 7, 1]. For example, the EC List [5] is based on a chemical-formula oriented classification. Thus it is sometimes difficult to classify the enzymes that catalyse several steps of reactions by creating intermediates. By contrast, mechanism-oriented approaches [1] classify enzymes based on reaction mechanisms such as (a) rules of the substructure changes in the chemical structures and (b) chemical structural reasons for the changes.

These classifications do not tend to take account of reaction-related issues such as inhibitors, pH, temperature, protein structure, amino-acid sequence, and the context of the metabolic network in which the reaction is taking place. A relational representation has the potential to capture many of these aspects simultaneously. We believe that such representations are mandatory if we hope to model biological systems from the micro to the macro-level in a seamless fashion.

In our experiments, the oxidoreductase and phosphotransferase classes are learned as logic programs. Oxidoreductases are enzymes which catalyse oxidisation and reduction. These reactions cause energy flow within organisms by exchange of electrons. Phosphotransferases are enzymes transferring a phosphate group from one compound (donor) to another (acceptor). The acceptors have electrophilic substructures such as NR, SR, and OR where N, S, O, and R are nitrogen, oxygen, sulphur, and alkyl group respectively.

Below we show where oxidoreductases and phosphotransferases fit within the EC classification system. The latest version of the EC List [5] contains 3196 enzymes, and is divided into 6 first-layers as follows:

1. Oxidoreductases (1st layer)
  1.1 Acting on the CH-OH group of donors (2nd layer)
    1.1.1 With NAD+ or NADP+ as acceptor (3rd layer)
      1.1.1.1 Alcohol dehydrogenase; Aldehyde reductase (4th layer)
        :
2. Transferases
        :
  2.7 Transferring Phosphorus-Containing Groups
        :
3. Hydrolases
4. Lyases
5. Isomerases
6. Ligases.

For example, the classification of EC enzyme Number 1.1.1.1 can be read as follows: EC Number 1.1.1.1 is an oxidoreductase which acts on the CH-OH group of donors, with NAD+ or NADP+ as acceptor, and the name of the enzyme is Alcohol dehydrogenase or Aldehyde reductase. Oxidoreductases are classified in EC Number 1.*.*.* and phosphotransferases are EC Number 2.7.*.*..

3 LIGAND Database

LIGAND is a database of chemical compounds and reactions in biological pathways [4]. The database consists of three sections: COMPOUND, REACTION, and ENZYME, and data is available in text files from the web site[9] and the anonymous ftp site[10]. The COMPOUND section is a collection of metabolic and other compounds such as substrates, products, inhibitors of metabolic pathways as well as drugs and xenobiotic chemicals. The REACTION section is a collection of chemical reactions involved in the pathway diagrams of the KEGG/PATHWAY database as well as in the ENZYME section. The ENZYME section is a collection of all known enzymatic reactions classified according to the EC List.

Knowledge integration could be performed for COMPOUND, REACTION, and ENZYME sections by cross-referring the EC numbers, the compound numbers, and the reaction numbers (Fig. 1). COMPOUND, REACTION, and ENZYME sections contain several atoms such as carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorous (P), sulphur (S), magnesium (Mg), manganese (Mn), iron (Fe) and iodine (I). Note that R is an alkyl group.

[9] http://www.genome.ad.jp/ligand/
[10] ftp.genome.ad.jp/pub/kegg/ligand/

REACTION

```
ENTRY       R00104
NAME        ATP:NAD+ 2'-phosphotransfera
DEFINITION  ATP + DAD+ <=> ADP + NADP+
EQUATION    C00002 + 2 C00003 <=> C00008 + C00006
PATHWAY     PATH: MAP00760 Nicotinate and nicotinamide metabolism
ENZYME      2.7.1.23
///
ENTRY       R00105
  :
```

COMPOUNDS

```
ENTRY       C00002
NAME        ATP
            Adenosine 5'-triphosphate
FORMULA     C10H16N5O13P3
REACTION    R00002 R00076 R00085 R00086 R00087 R00088 R00089 R00104
  :
///
ENTRY       C00003
NAME        NAD
FORMULA     C21H28N7O14P2
  :
///
ENTRY       C00006
NAME        NADP
FORMULA     C21H29N7O17P3
  :
///
ENTRY       C00008
NAME        ADP
FORMULA     C10H15N5O10P2
  :
```

Fig. 1. Excerpt from REACTION and COMPOUND sections in LIGAND

4 Experiments

The experiments in this section are aimed at evaluating the following null hypothesis:

Null hypothesis: A relational representation cannot capture enzyme classification rules based only on descriptions of the underlying biochemical reactions.

4.1 Materials

The ILP system used in the experiments is Progol 4.4 [11]. In order to allow reproducibility of the results, the data sets and Progol's settings used in the experiments have also been made available [12].

Our study is restricted to learning classification rules for two classes of enzymes: the main class EC1 (Oxidoreductase) and a more specific class EC2.7 (Phosphotransferase). One justification for these choices is that EC1 and EC2 are the most populated classes of enzymes and they contain a relatively large number of examples, which means that learning is more robust and the results more meaningful. Other classes contain a smaller number of enzymes, for example EC5 and EC6 each contain around 200 known enzymes compared to EC1 and EC2 with over 1000 enzyme in each.

4.2 Methods

For the experiments reported in this section, we use a relational representation to represent the biochemical reactions catalysed by each enzyme. In this representation, each reaction is defined as a set of compounds in the left hand side (LHS) and the right hand side (RHS) of the reaction. For example, the enzyme with EC number 1.1.1.37 which belongs to the class of Oxidoreductase and catalyses the reaction $C00149 + C00003 \rightleftharpoons C00036 + C00004 + C00080$ is represented by the following Prolog facts:

```
oxidoreductase('1.1.1.37').
lhs('1.1.1.37','C00149').
lhs('1.1.1.37','C00003').
rhs('1.1.1.37','C00036').
rhs('1.1.1.37','C00004').
rhs('1.1.1.37','C00080').
```

For each chemical compound, we only represent the number of atoms of each element appearing in the compound. For example, compound $C00003$ with chemical formula $C_{21}H_{28}N_7O_{14}P_2$ can be represented as follows.

[11] Available from: http://www.doc.ic.ac.uk/~shm/Software/progol4.4/
[12] Available from: http://www.doc.ic.ac.uk/bioinformatics/datasets/enzymes/

```
compound('C00003').
atoms('C00003','c',21).
atoms('C00003','h',28).
atoms('C00003','n',7).
atoms('C00003','o',14).
atoms('C00003','p',2).
```

In order to capture the exchange of elements in compounds during the reactions, we define the relation 'diff_atoms/5' which represents the difference between the number of particular atoms in compounds $C1$ and $C2$ which appeared in LHS and RHS respectively.

```
diff_atoms(Enz,C1,C2,E,Dif):-
     lhs(Enz,C1),
     rhs(Enz,C2),
     atoms(C1,E,N1),
     atoms(C2,E,N2),
     Dif is N1 - N2, Dif > 0 .
```

We report on two series of experiments for each of Oxidoreductase and Phosphotransferase enzyme classes. In 'Mode 1', the hypotheses language was limited so that only 'diff_atoms/5' and numerical constraint predicates (i.e. $=$, $\geq$ and $\leq$) can appear in the body of each hypothesis. In 'Mode 2', the hypothesis language also included 'atoms/3', 'lhs/2' and 'rhs/2' (Mode 2). Part of the mode declaration and background knowledge used by Progol are shown in Table 1.

In the experiments, we compared the performance of Progol in learning classification rules for each of Oxidoreductase and Phosphotransferase enzyme classes from varying-sized training sets. Figure 2 shows the experimental method used for this purpose. The average predictive accuracy was measured in 20 different runs. In each run, the number of positive and negative training examples was varied while the number of 'hold-out' test examples was kept fixed (i.e. 200). Test-set positive examples were randomly sampled from the target enzyme class. Negative examples for the target class EC1 (Oxidoreductase) were randomly sampled from other major classes (i.e. EC2 to EC6). For EC2.7 (phosphotransferase), negative examples were randomly sampled from other sub-classes of EC2. Progol was then run on the training examples using 'Mode 1' and 'Mode 2'. For each iteration of the loop the predictive accuracy of the learned classification rule was measured on the test examples. The average and standard error of these parameters were then plotted against the number of training examples.

4.3 Results

The results of the experiments are shown in Figure 3. In these graphs, the vertical axis shows predictive accuracy and the horizontal axis shows the number of training examples. For each experiment, predictive accuracies were averaged over 20 different runs (error bars represent standard errors). According to these graphs,

Table 1. Part of mode declarations and background knowledge used by Progol in the experiments.

```
:- set(h,10000)?
:- set(r,100000)?
:- set(noise,5)?

:- modeh(1,oxidoreductase(+enzyme))?
:- modeb(*,diff_atoms(+enzyme,-compound,-compound,#element,-int))?
:- modeb(1,eq(+int,#int))?
:- modeb(1,lteq(+int,#int))?
:- modeb(1,gteq(+int,#int))?

% The following mode declarations were added in 'Mode 2' experiments
%:- modeb(*,lhs(+enzyme,-compound))?
%:- modeb(*,rhs(+enzyme,-compound))?
%:- modeb(*,atoms(+compound,#element,-nat))?

element(c). element(h). element(n). element(o). element(p).
element(s). element(mg). element(mn). element(fe). element(i).

diff_atoms(Enz,C1,C2,E,Dif):-
       lhs(Enz,C1),
        rhs(Enz,C2),
        atoms(C1,E,N1),
        atoms(C2,E,N2),
        Dif is N1 - N2, Dif > 0 .

eq(X,X):-
        not(var(X)),
int(X),!.

gteq(X,Y):-
        not(var(X)), not(var(Y)),
        int(X), int(Y),
        X >= Y, !.
gteq(X,X):-
        not(var(X)),
        int(X).

lteq(X,Y):-
        not(var(X)), not(var(Y)),
        int(X), int(Y),
        X =< Y, !.
lteq(X,X):-
        not(var(X)),
        int(X).
```

for i=1 **to** 20 **do**
  for j **in** (10,20,40,80,160) **do**
    Randomly sample j positive and j negative 'training' examples
    Randomly sample 200 positive and 200 negative 'test' examples
    Run Progol on the 'training' set using 'Mode 1'
      A_{ij} =predictive accuracy of the learned classification rule on the 'test' set
    Run Progol on the 'training' set using 'Mode 2'
      A'_{ij} =predictive accuracy of the learned classification rule on the 'test' set
  end
end
for j **in** (10,20,40,80,160) **do**
  Plot average and standard error of A_{ij} and A'_{ij} versus j ($i \in [1..20]$)

Fig. 2. Experimental method.

the overall predictive accuracies of the learned rules for the Phosphotransferase dataset are higher than the overall predictive accuracies for the Oxidoreductase dataset. These results suggest that using only 'diff_atoms/5' information is sufficient to get a relatively high accuracy for the phosphotransferase dataset while this is not the case for Oxidoreductase dataset and probably we require additional information (e.g. knowledge about the structure) which cannot be captured by 'diff_atoms/5'. For the Phosphotransferase dataset the accuracy difference between using mode declarations 'Mode 1' and 'Mode 2' is not significant, however, for the Oxidoreductase dataset 'Mode 2' clearly outperforms 'Mode 1' in all experiments. In both graphs the null hypothesis is refuted as the predictive accuracies are significantly higher than default accuracy (i.e. 50%).

In the next section we discuss some of the descriptions which have been learned for each of the target enzyme classes.

5 Discussion

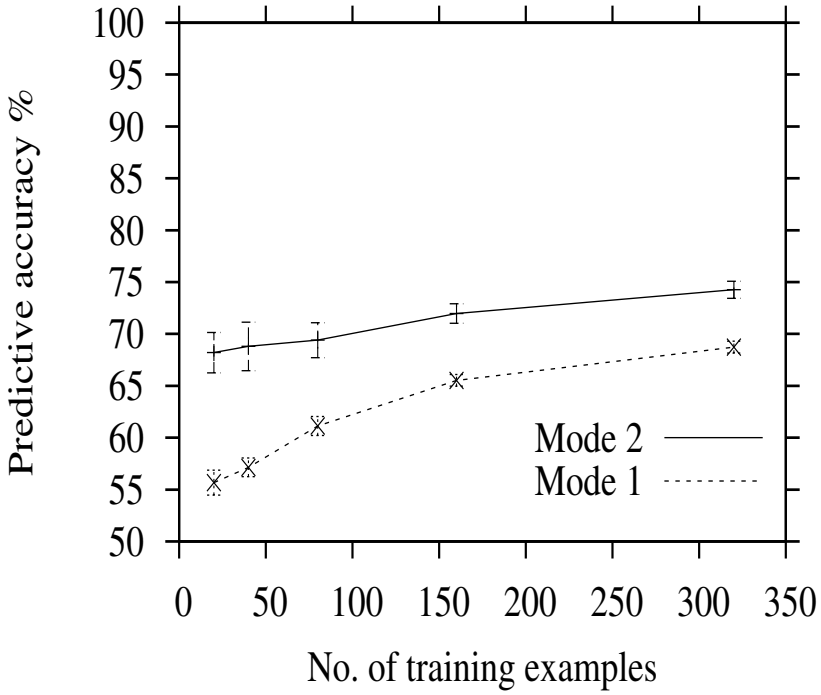
In this paper we have made an initial attempt at machine learning EC classes for enzyme function prediction based on biochemical reaction descriptions found in the LIGAND database. Figure 4 shows a diagrammatic representation of the class of chemical reaction catalysed by Oxireductase and Phosphotransferase enzymes. We succeeded in learning descriptions of the Oxidoreductase and Phosphotransferase class from LIGAND database in the form of a logic program containing the following rules (among others).

Oxidoreductases Rule

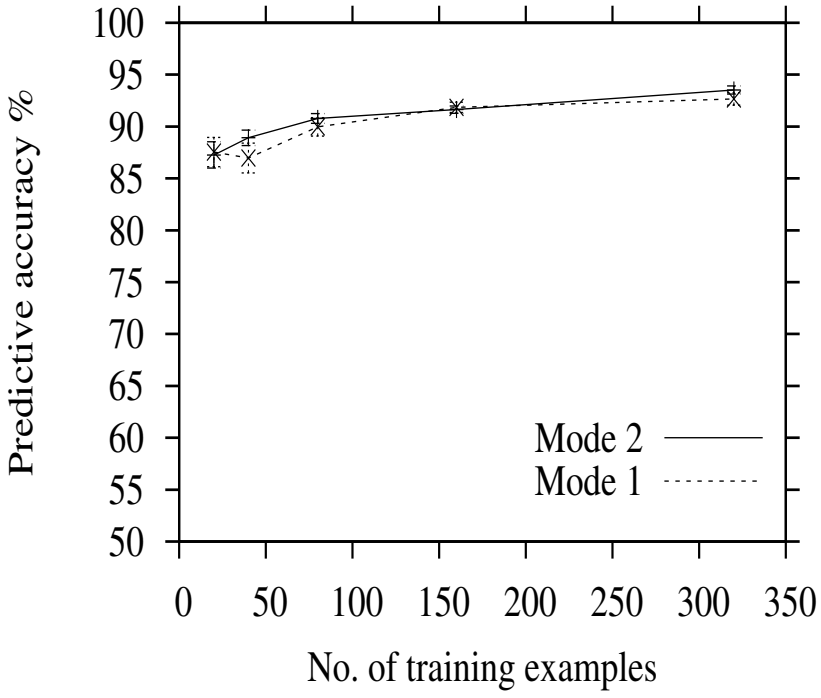
```
oxidoreductase(A) :- diff_atoms(A,B,C,h,D), atoms(B,o,E),
                     atoms(C,o,E), eq(D,2), lteq(E,4).
```

Phosphotransferases Rule

```
phosphotrans(A) :- diff_atoms(A,B,C,h,D), diff_atoms(A,B,C,o,E),
```



(a) Oxidoreductase



(b) Phosphotransferase

Fig. 3. Performance of Progol in learning enzyme classification rules for a) Oxidoreductase and b) Phosphotransferase. In both graphs default accuracy is 50%.

diff_atoms(A,B,C,p,F), eq(F,1), lteq(E,7).

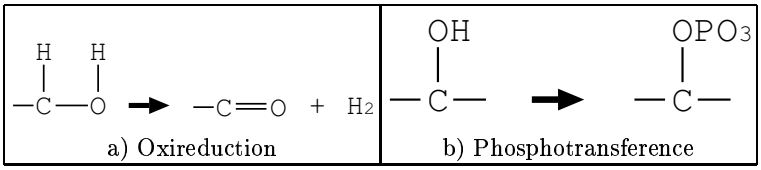


Fig. 4. Chemical reactions catalysed by Oxireductase and Phosphotransferase enzymes

In the above the Oxidoreductase Rule captures the elimination of H_2 which is typical of oxidation-reduction reactions (see Figure 4a). The Phosphotransferase Rule represents the exchange of the phosphate group PO_3 (Figure 4b). Logical speaking, the boundary constraint `lteq(E,7)` is consistent with a transfer of three Oxagen atoms. However, it is not clear why `lteq(E,7)` is learned instead of `eq(E,3)`. Further analysis by domain experts is required to identify the chemical meaning of this constraint.

6 Conclusion

As mentioned previously, resources such as the Gene Ontology (GO) [3] and the Enzyme Classification (EC) [5] system both provide isa-hierarchies of enzyme functions. On the face of it GO and EC should be invaluable for supporting automation within Functional Genomics, which aims at predicting the function of unassigned enzymes from the genome projects. However, neither GO nor EC can be directly used for this purpose since the classes presently have only a natural language description. The study described in this paper has taken a first step toward automatic formulation of rules which describe some of the major functional classes of enzymes. By extending this study we believe it should be possible to learn descriptions for all major GO and EC classes. However, in order to do so we will need to involve domain experts to check the quality and comprehensibility of the learned rules.

In order to speed-up the learning process in this study, we simply compared the number of atoms between two compounds with `diff_atoms/5` predicate. The limitation of this representation is that we ignore the structure of compounds. For example, enzymes which catalyse the elimination of H_2 are called dehydrogenase, and the reaction results in a double bond between C-O, C-C, or C-N. By considering the types of bonds between atoms such as single bond and double bond, we could track the introductions of double bonds between atoms and determine the locations where H_2 is eliminated. Logic programs could represent the structural information by expressing connections and the type of connections between atoms.

The learned knowledge could be viewed as not only rules for classification but also programs for a logic-based biological simulation. As a future study, we believe it would be worth adding more background knowledge including inhibitors, cofactors and amino-acid sequential information which is available from various public-domain biological databases.

Acknowledgements

This work was supported by the ESPRIT IST project "Application of Probabilistic Inductive Logic Programming (APRIL)", the BBSRC/EPSRC Bioinformatics and E-Science Programme, "Studying Biochemical networks using probabilistic knowledge discovery" and the DTI Metalog project.

References

1. M. Arita and T. Nishioka. Hierarchical classification of chemical reactions. *Bio Industry*, 17(7):45–50, 2000.
2. C.H. Bryant, S.H. Muggleton, S.G. Oliver, D.B. Kell, P. Reiser, and R.D. King. Combining inductive logic programming, active learning and robotics to discover the function of genes. *Electronic Transactions in Artificial Intelligence*, 5-B1(012):1–36, November 2001.
3. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
4. S. Goto, Y. Okuno, M. Hattori, T. Nishioka, , and M. Kanehisa. Ligand: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Research*, 30:402–404, 2002.
5. International Union of Biochemistry and Molecular Biology. *Enzyme Nomenclature: Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, New York, 1992.
6. M. Turcotte, S.H. Muggleton, and M.J.E. Sternberg. Automated discovery of structural signatures of protein fold and function. *Journal of Molecular Biology*, 306:591–605, 2001.
7. C. Walsh. *Enzymatic Reaction Mechanisms*. W. H. Freeman and Company, 1979.
8. M. R. Wilkins, K. L. Williams, R. D. Appel, and D.F. Hochstrasser. *Proteome Research : New Frontiers in Functional Genomics (Principles and Practice)*. Springer Verlag, Berlin, 1997.