# Identification of Divergent Functions in Homologous Proteins by Induction over Conserved Modules

**Imran Shah**
American Type Culture Collection
12221 Parklawn Drive
Rockville, MD 20852
Tel: (301) 881-2600
Fax: (301) 816-4379
Email: ishah@atcc.org

**Lawrence Hunter**
Bldg. 38A, 9th fl, MS-54
National Library of Medicine
Bethesda, MD 20894 USA
Tel: (301) 496-9303
Fax: (301) 496-0673
hunter@nlm.nih.gov

### Abstract

Homologous proteins do not necessarily exhibit identical biochemical function. Despite this fact, local or global sequence similarity is widely used as an indication of functional identity. Of the 1327 Enzyme Commission defined functional classes with more than one annotated example in the sequence databases, similarity scores alone are inadequate in 251 (19%) of the cases. We test the hypothesis that conserved domains, as defined in the ProDom database, can be used to discriminate between alternative functions for homologous proteins in these cases. Using machine learning methods, we were able to induce correct discriminators for more than half of these 251 challenging functional classes. These results show that the combination of modular representations of proteins with sequence similarity improves the ability to infer function from sequence over similarity scores alone.
**Keywords:** protein function; protein sequence; protein modules; protein function; Enzyme Commission; representation; machine learning

## Introduction

Homologous proteins do not necessarily exhibit identical biochemical function; in fact, functional divergence is required for organismal evolution. Despite this caveat, local or global sequence similarity between proteins is widely used as an indication of functional identity. In this paper, we describe the use of supervised machine learning systems to induce sequence-based methods for identification of divergent functions in homologous proteins.

### Background

Recently, we systematically assessed the reliability of function imputation by pair-wise sequence alignment, using the Enzyme Commission (EC) classification as our definition of function (Shah & Hunter 1997). In that work, we found that sequence similarity scores are not sufficient to accurately assign protein sequences to many EC functional classes, no matter whether gapped or ungapped alignments are used, and regardless of where similarity score thresholds may be set.



The most common problem we found in mapping from sequence to functional class is that of false positives; that is, sequences which are similar to a query, but have a different functional class. The dependence of enzymatic function on the fine details of three-dimensional protein structure and associated chemistry suggests that this problem may be intractable in general. However, inspection of specific errors suggested to us that in some cases it may be possible to automatically identify regions of a sequence that must be conserved in order that function also be conserved. This observation is compatible with theories of modular protein evolution, e.g. (Doolittle & Bork 1993). On the basis of this observation, we designed a series of machine learning experiments to test the hypothesis.

### Strategy

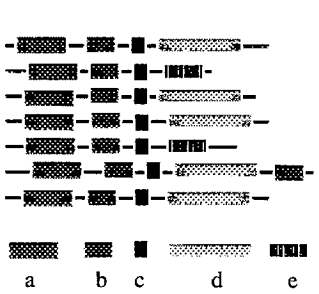
Our approach is to use supervised machine learning to identify protein modules which can be used to discriminate among EC functional classes of similar proteins. First, we identified the set of proteins that were the target of our study: those that had significant sequence similarity to at least one protein of another EC class. Then we developed a simple vector representation of the modules present in each protein, based on the ProDom database (Sonnhammer & Kahn 1994). We applied both statistical and information theoretic induction methods to these representations, and generated unbiased estimates of the ability of the induced classifiers to assign the correct EC class to similar sequences, based on the modular structure of the proteins.

## Methods

### Definition of training sets

Our universe of sequences was defined by taking all of the sequences from Swiss-Prot release 33 (Bairoch & Boeckmann 1992) that were labeled with an EC classification from Enzyme release 21 (Bairoch 1994). [1]

[1] We used the EC classification as a gold standard for protein function. We recognize that this classification is flawed in various ways, but because of its breadth and the dearth of reasonable alternatives, we feel it is appropriate




(i) conserved modules

| a | b | c | d | e | b' |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 0 | 0 |

(ii) representation

Figure 1: Representing protein sequences on conserved modules. (i) A set of (hypothetical) homologous proteins with conserved modules shown as boxes. There are a total of five modules: a,b,c,d and e. (ii) A representation of the proteins on the basis of the modular attributes. Each protein is shown as a vector of attribute values. Modules which occur more than once in a sequence are called repeats. The second occurrence of module b is treated as a separate attribute, b'.

Note that not all proteins in an EC class must be homologous. Non-homologous subgroups within EC classes may arise due to different evolutionary origins of enzyme subunits, convergent evolution of proteins catalyzing a particular reaction, or by vague or generalized reaction definitions by the EC. Since we are concerned with detecting proteins with similar sequences but divergent functions, we needed to control for the presence of non-homologous proteins in EC classes. We did so by subdividing the EC classes with non-homologous subgroups on the basis of sequence similarity. We call these homologous subsets of EC classes *simgroups*. If an EC class has no non-homologous sequences, all of the sequences in that class are assigned to a single simgroup. All members of any simgroup have the same (EC) function and similar sequences.

We defined sets of positive and negative examples for learning from each simgroup. The positive examples were taken to be all the members of a simgroup. The negative examples were the union of all proteins in our universe that had any significant sequence similarity to any member of the simgroup, but were not themselves members of the group. These proteins are probably homologous to the members of the simgroup, but exhibit different functions. We generate training sets of positive and negative examples for each simgroup. The goal of learning is to be able to discriminate among these positive and negative examples for each set.

Our universe contained the 15,208 SwissProt proteins (out of 52,205 total) which are labeled with one of 1,327 EC classes. From this universe, we were able to define 251 training sets (simgroups and homologous proteins with different functions) containing at least ten or more positive examples, and at least one negative example.

## Representing Proteins

Homologous proteins with divergent functions may exhibit differences at many levels of description, ranging from point mutations (say, in an active site) to large-scale rearrangements. Unfortunately, it is not possible to test all possible descriptions, since the number of training examples for most functional classes is small, and learning algorithms require a number of training example proportional to the number of possible discriminators tested. For this reason, we had to carefully identify types of descriptions that we felt had a reasonable chance of being able to make the discriminations *a priori*.

Based on modular theories of protein evolution, and on our observations in the systematic study, we decided to test if the presence and arrangement of conserved subregions of sequence ("domains") could be used to discriminate among functions. Although using a multiple sequence alignment to identify differences at the level of individual amino acids might be desirable, there are far too many differences at this level of description for effective induction.

The ProDom database (Sonnhammer & Kahn 1994) is one attempt to systematically define protein domains, done on the basis of local sequence alignments within a large set of sequences. Since all of our training sets consisted entirely of homologous proteins, we were able to define a simple attribute vector identifying the presence or absence of all domains that are observed anywhere in the training set (figure 1).

## Induction

We tested two approaches to inducing discriminators able to segregate the positive from the negative

to use it for this study.

| L | P | Accuracy range | | | |
|---|---|---|---|---|---|
| | | <0.9 | ≥0.9 | ≥0.95 | 1 |
| NB | $P^+$ | 71 (28%) | 180 (72%) | 155 (62%) | 140 (56%) |
| | $P^-$ | 8 (3%) | 243 (97%) | 225 (90%) | 160 (64%) |
| | $P_t$ | 11 (4%) | 240 (96%) | 223 (89%) | 132 (53%) |
| C4.5 | $P^+$ | 85 (34%) | 166 (66%) | 145 (58%) | 129 (51%) |
| | $P^-$ | 1 (0%) | 250 (100%) | 236 (94%) | 177 (71%) |
| | $P_t$ | 10 (4%) | 241 (96%) | 227 (90%) | 134 (53%) |

Table 1: Summary of learning performance for 251 data sets. The performance results are summarized by learners from top to bottom, and accuracy range from left to right. The first column, labeled **L** shows the learners: naive Bayes (NB) and C4.5. The next column shows performance (**P**): $P_t$ (total), $P^+$ (same function), and $P^-$ (different function). Each cell shows the number of datasets with performance P, for a given measure and learner, and the percentage of datasets that represents.

examples: information-theoretic decision trees and naive Bayesian discrimination (NB). The information theoretic method we used, C4.5 (Quinlan 1993), recursively looks for an attribute (in this case, presence or absence of a particular domain) which allows the training examples to be divided such that there is an information gain (Shannon 1948) after the split. This method is widely used, and makes minimal assumptions about the data. The alternative method, naive Bayes (e.g. see (Mitchell 1997)) makes much stronger assumptions about the data, namely that each attribute is statistically independent of all of the others. Under that assumption, NB provides the best possible discriminator. When the assumption is not true (as it is clearly not in this case), NB may nevertheless generate a powerful discriminator. Although many other supervised induction methods exist (e.g. artificial neural networks), these two methods represent quite different approaches, so that if there were going to be differences in performance attributable to induction method, it is likely to be apparent here. Also, these two methods are very quick to train, which is important given our 251 datasets. Our experience suggests that neural networks would perform at about the same level as these two methods, and would have taken much longer to train.

We used 10-way cross-validation to generate unbiased estimates of the accuracy of both of these methods on each of our 251 training sets.

## Results

Table 1 summarizes the results of induction on all 251 simgroups. Using either method, it is possible to use ProDom domains to perfectly discriminate between homologs with similar functions and homologs whose function has diverged in more than half of the simgroups. In more than two thirds of the simgroups, it is possible to identify both same-function and different-function sequences at greater than 90% accuracy.

Also note that, although there were some differences between the induction methods in sensitivity and specificity (P+ and P- in Table 1), overall accuracy (Pt in Table 1) was remarkably similar for both methods.

Results for some representative simgroups are discussed below.

### Short Chain Alcohol Dehydrogenases

The short chain alcohol dehydrogenases (SC-ADH) data set consists of 210 examples all from the EC sub-sub-class 1.1.1.*. The forty-five positive examples in this set are from a single simgroup within EC 1.1.1.1 (this class has two non-homologous simgroups) and the negative examples are all from other child nodes of EC 1.1.1.*. Because the functions are closely related, this is a somewhat difficult test. By visual inspection of the conserved regions in this data set (see figure 2), it appears that EC 1.1.1.1 members can be distinguished from others on the basis of modules 238 and 237. The negative instances possessing modules 238 and 237 also contain module 4870, which is not present in the positive examples. Absent from the positive class and present in the negative is module 36. The induction methods made varying use of these observations.

The conditional probabilities computed by NB are shown in table 2. The individual conditional probabilities of the positive and negative classes, for given feature values, contribute towards the calculation of the posterior probability. The binary feature values in this case convey the presence or absence of a module. From row 3 it can be seen that the absence of modules 36, 201 and 4870, and presence of modules 12, 237 and 238 are good predictors of the positive instances. The converse is true for the negative instances.

The information theoretic test used in constructing the decision tree gives a more succinct hypothesis for discriminating positive and negative instances. The decision tree in figure predicts the positive class if

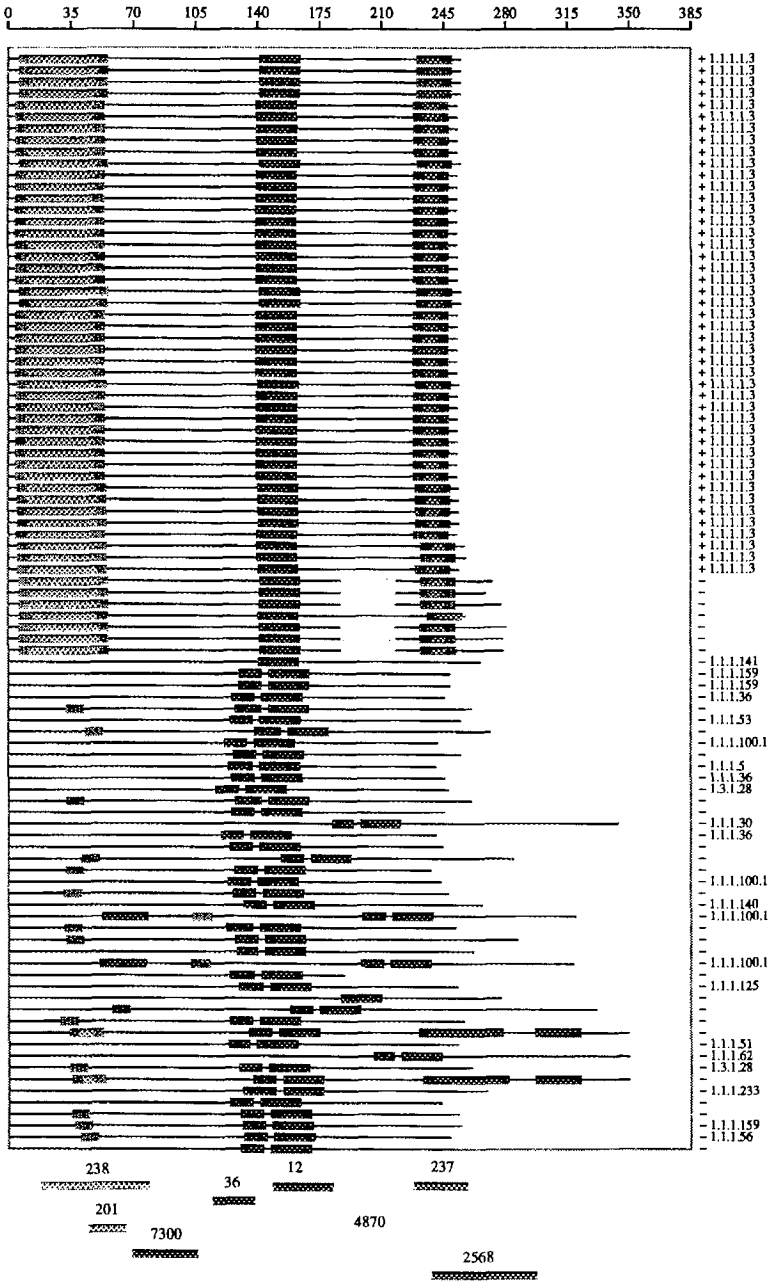



Figure 2: The figure shows conserved modules in the examples in data set EC 1.1.1.1(3). Each protein sequence is represented by line, and conserved modules as rectangular boxes of the same color. Each example is labeled on the right with + or − and EC number, showing class membership. The similarity between the examples decreases from top to bottom. A scale on the top of the figure shows the approximate length of the proteins; a legend at the bottom shows the ProDom ID of the modules.

Marginal probability of same function $P(+) = 0.474$
Marginal probability of different function $P(-) = 0.526$

| *Module* | 36 | 12 | 201 | 4870 | 237 | 238 | 36 | 201 |
|---|---|---|---|---|---|---|---|---|
| *Value* | - | + | - | - | + | + | + | + |
| $LP_c^+$ | 0.778 | 0.778 | 0.778 | 0.778 | 0.778 | 0.778 | 8.444 | 8.499 |
| $LP_c^-$ | 1.632 | 1.632 | 1.523 | 0.972 | 0.834 | 0.778 | 1.222 | 1.699 |

Table 2: The probabilities used by the naive Bayesian classifier for EC 1.1.1.1. From top to bottom, the columns show: the ProDom module ID; its value, present (+) or absent (-); negative logarithm of the conditional probability that the function is the same, given the value $LP_c^+ = -log_{10}(P(+|x = v))$; and negative logarithm of the conditional probability the function is different, given the value $LP_c^- = -log_{10}(P(-|x = v))$.

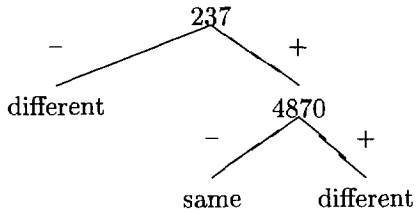



Figure 3: The decision tree for dataset EC 1.1.1.1 (simgroup 3). The numbers at the nodes refer to ProDom modules. The branches are labeled '+' for module present and '–' for module absent. The leaves are labeled 'same' if the function is the same, and 'different' if the function is different.

module 237 (or 238) is present and 4870 absent. This tree is nearly correct, with the exception of one false positive. This discriminator corresponds well with our intuitions from visual inspection of the data set (figure 2).

In this case, both induction methods were capable of finding discriminators which successfully identified functional differences among homologs. Although visual inspection suggests that this was not a particularly difficult example, it does help validate the strategy.

## Dihydrofolate reductase

Proteins which have multiple catalytic domains pose a challenge to function prediction methods which use sequence similarity alone. It is reasonable to expect that in these cases the use of information about conserved regions will greatly aid in discrimination. Dihydrofolate reductase (EC 1.5.1.3) and thymidylate synthase (EC 2.1.1.45) activities were found to occur in multidomain enzymes with significant sequence similarity. In this example, both induction methods were able to find perfect discriminators.

Figure 4 shows conserved modules in the single and multidomain examples of the enzymes. Module 375 only occurs in dihydrofolate reductase, while thymidylate synthase has module 504. The multidomain examples have dihydrofolate reductase activity in the N-terminal and thymidylate synthase in the C-terminal. The conditional probabilities for the binary modular features in table agree with this. The first two columns show a high probability for the presence and absence of modules 375 and 504, respectively.

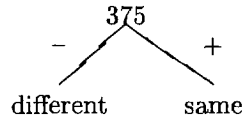



Figure 5: Decision tree for data set EC 1.5.1.3(1)

The decision tree (figure 5) embodies the simplest hypothesis for discrimination. Hence, the presence of module 375 predicts EC 1.5.1.3(1) and its absence, EC 2.1.1.45(1).

## L-lactate dehydrogenase

L-lactate dehydrogenase (EC 1.1.1.27) poses a somewhat greater challenge for our method. False positive matches occur with malate dehydrogenase, which is functionally quite similar to lactate dehydrogenase. Both enzymes contain a central conserved region (module 139, figure 6) but the remaining regions differ to varying degrees. Although both induction methods did reasonably well, two negative examples were not differentiated by either discriminator. One of these instances has only been assigned partial EC classification 1.1.1 (DHL2_LACCO) and may in fact possess lactate dehydrogenase activity. The other is an archaebacterial malate dehydrogenase (MDH_HALMA), which genuinely cannot be distinguished on the basis of domain structure.

# Discussion and Conclusions

These results demonstrate that the presence or absence of particular ProDom modules can often be used to discriminate among functionally distinct homologs. The induced discriminators can be used to detect situations in which sequence similarity alone may provide misleading suggestions of protein function, and

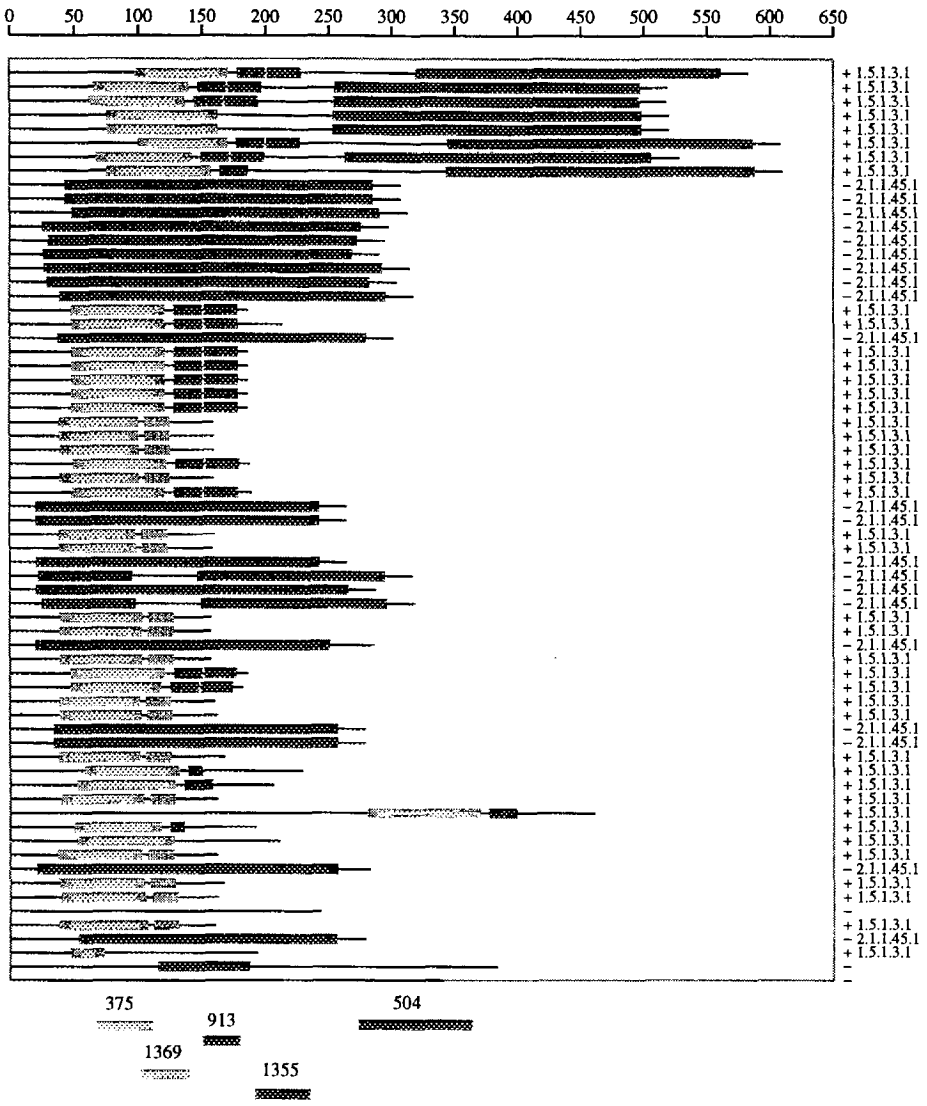



Figure 4: Conserved modules in EC 1.5.1.3

Marginal probability of same function $P(+) = 0.636$
Marginal probability of different function $P(-) = 0.364$

| *Module* | 375 | 504 | 1355 | 1369 | 913 | 913 | 1369 | 1355 |
|---|---|---|---|---|---|---|---|---|
| *value* | + | - | - | - | + | - | + | + |
| $LP_c^+$ | 0.699 | 0.791 | 0.907 | 0.924 | 1.000 | 1.000 | 1.092 | 1.118 |
| $LP_c^-$ | 8.492 | 1.778 | 0.699 | 0.699 | 8.589 | 0.699 | 8.714 | 8.647 |

Table 3: Probabilities for computing naive Bayesian for EC 1.5.1.3(1)

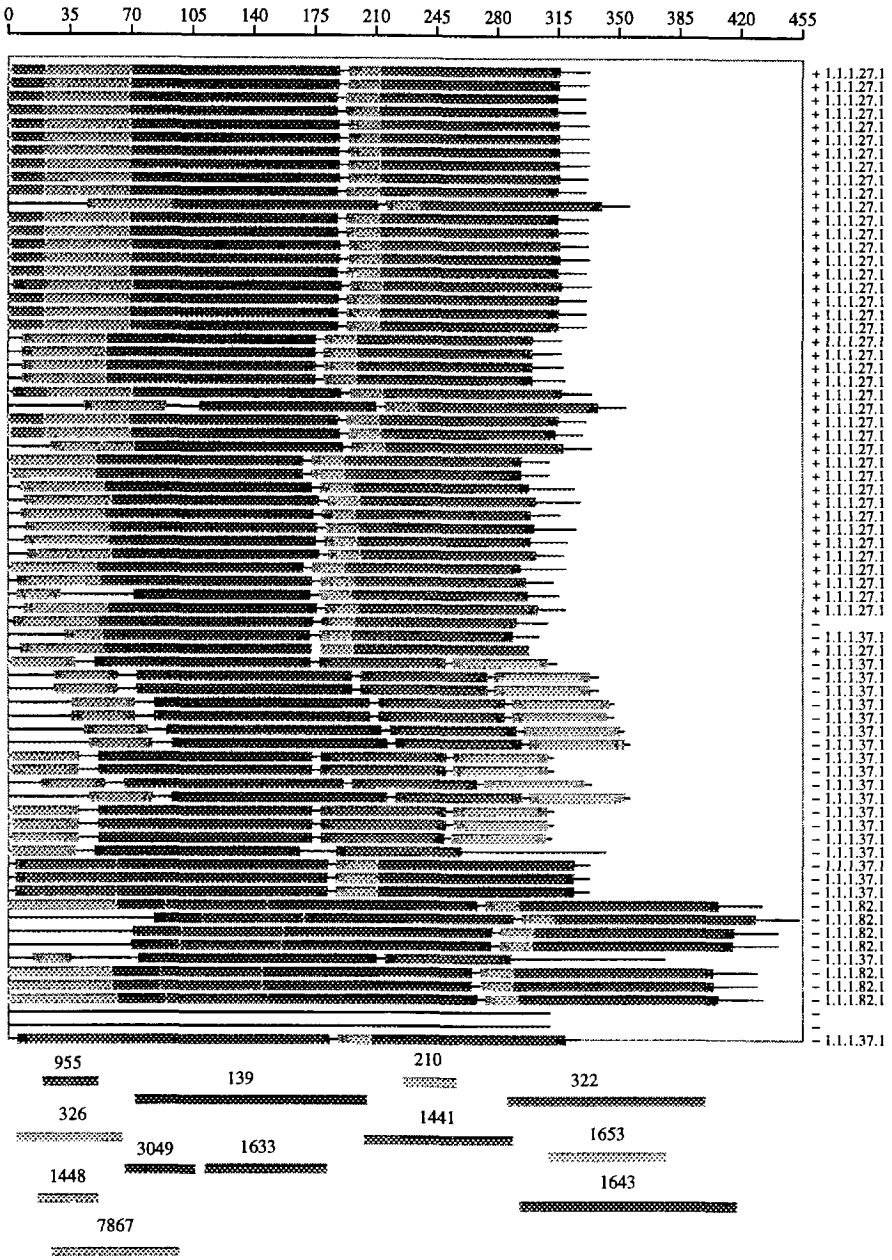



Figure 6: Conserved modules in EC 1.1.1.27(1)

Marginal probability of same function $P(+) = 0.575$
Marginal probability of different function $P(-) = 0.425$

| *Module* | 322 | 326 | 210 | 1448 | 1441 | 1653 | 1633 | 1643 |
|---|---|---|---|---|---|---|---|---|
| *value* | + | + | + | - | - | - | - | - |
| $LP_c^+$ | 1.041 | 1.041 | 1.041 | 1.041 | 1.041 | 1.041 | 1.041 | 1.041 |
| $LP_c^-$ | 2.232 | 2.232 | 1.419 | 1.357 | 1.357 | 1.302 | 1.232 | 1.232 |

Table 4: Probabilities for computing naive Bayesian for EC 1.1.1.27(1)

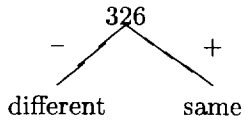



Figure 7: Decision tree for data set EC 1.1.1.27

may therefore be useful in general sequence analysis.

However, there remain a large number of cases in which even the combination of sequence similarity and ProDom domains does not appear to be adequate to induce discriminators among alternative functional possibilities. This may be due to a number of factors.

First, the representation used in this work depends on the assumptions underlying the construction of the ProDom database. ProDom is constructed using ungapped pair-wise sequence alignments and an iterative strategy for identifying boundaries of conserved regions. It is possible that an alternative representation of domains might make possible the induction of more accurate discriminators.

Second, differences in catalytic activity can result from sequence changes as modest as a single amino acid. Domain level representations are not sufficiently expressive to capture these differences. However, more expressive representations must be chosen with caution. Richer representations admit richer hypothesis spaces, and the number of available examples may not be adequate to select good discriminators among so many possibilities.

Third, the definitions from the EC that allow us to identify homologous proteins with different functions may be flawed. The mechanism by which the EC divides biochemical functions into classes, and the mechanism by which individual proteins are assigned to those classes are both subject to human error. Furthermore, it may be the case that, especially in eukaryotes, there simply is not a one-to-one mapping between proteins and functions; e.g., an individual protein may catalyze multiple reactions. Any errors in assigning functions to proteins will make perfect discrimination impossible.

Fourth, it may be the case that alternative induction approaches may do significantly better than the two that we tested. Since our methods are so different from each other, we doubt that the application of, for example, neural networks to the representation we used will significantly improve total performance. However, since NB did have better sensitivity than C4.5 and C4.5 had better specificity, it may be the case that a stacking or voting scheme may produce a better combined classifier than either alone.

In conclusion, we have demonstrated that it is possible to improve on the performance of simple similarity scores in assigning function to protein sequences. It is also clear that methods better than ours remain to be discovered.

## References

Bairoch, A., and Boeckmann, B. 1992. The SWISS-PROT protein sequence data bank. *Nucleic Acids Research* 20:2019–2022.

Bairoch, A. 1994. The ENZYME data bank. *Nucleic Acids Res.* 22:3626–3627.

Doolittle, R., and Bork, P. 1993. Evolutionarily mobile modules in proteins. *Sci Am* 269(4):50–56.

Mitchell, T. 1997. *Machine Learning.* McGraw Hill.

Quinlan, J. 1993. *C4.5: Programs for Machine Learning.* Morgan Kauffman.

Shah, I., and Hunter, L. 1997. Predicting enzyme function from sequence: A systematic appraisal. *ISMB* 5:276–283.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Jounal* XXVII(3):379–423.

Sonnhammer, E., and Kahn, D. 1994. The modular arrangement of proteins as inferred from analysis of homology. *Protein Science* 3:482–489.