Visualization based on the Enzyme Commission nomenclature

I. Shah
Computational Sciences and Informatics
George Mason University
Fairfax, VA 22030 USA
ishah@gmu.edu

L. Hunter
Bldg. 38A, 9th fl, MS-54
National Library of Medicine
Bethesda, MD 20894 USA
hunter@nlm.nih.gov

We developed a tool for visualizing data related to protein function, based on the nomenclature defined by the International Enzyme Commission. The method represents the 1327 specific reaction types and three additional levels of abstract classifications in the nomenclature as an interactive graph. Each node and link in the graph can have associated with it stored or computed data values of various types, each of which can be assigned to any of a set of visualization methods, including color, size, various chart types and text. Visualizations can then be interactively created for single nodes, groups of nodes or the entire graph. This visualization tool is particularly useful for exploring the distribution of quantitative attributes across protein functional classes. As a test of this tool, we developed a visualization of data measuring the effectiveness of sequence similarity for prediction of EC class [1]. Using this tool, it was possible to rapidly scan hundreds of functional classes, and to correlate predictability from sequence with other attributes, facilitating the generation of hypotheses about the causes of predictive failure.

1 Introduction

Enzyme function is the cornerstone of much of biology. Using the tree-structured Enzyme Commission (EC) nomenclature, we have developed a system for graphically visualizing various kinds of data based on its relationship to specific enzyme functions. The tool facilitates navigation through the EC hierarchy and the visualization of coherent subsets of enzyme functions. Each node in the enzyme function graph can be associated with stored or computed data about that functional class, and such data can be visualized using color, size, distance and various other mechanisms. In addition, each node in the graph is active, allowing the invocation of a variety of other tools on any particular portion of the function hierarchy.

We used this visualization tool to explore the predictability of enzyme function from sequence. Identification of enzymatic function from protein se-

EC	Name	nodes at level		
		2	3	4
1	Oxidoreductases	19	64	311
2	Transferases	9	24	333
3	Hydrolases	8	42	430
4	Lyases	7	15	131
5	Isomerases	6	15	54
6	Ligases	5	9	68
	Total	54	169	1327

Table 1: Summary of nodes at different levels of the EC classification. Rows corresponds to a top level classification. Columns 3,4 and 5 show the number of nodes at each level. For example, the oxidoreductases contain 19 second level groups, 64 third level groups and 311 leaves. Only those EC classes with at least one protein are included in this table.

quence is becoming increasingly important, particularly as the availability of data about whole genomes grows at a rapid rate. Sequence comparison is one of the most widely used methods for predicting protein function. However, a systematic study of the use of sequence similarity for predicting enzyme function [1] demonstrates that this method is often fallible. However, there is wide variability in the reliability of such predictions among different classes of enzyme function. The first application of our visualization system, described in detail in this paper, has been to explore the reliability of functional prediction from sequence.

2 Implementation

The design of the application is based on the EC classification [2] [3]. EC classes define enzyme function based on the reaction which is catalyzed by the enzyme. The classification scheme is hierarchical, with four levels. There are six broad categories of function at the top of this hierarchy and about 3,500 specific reaction types at the bottom. EC classes are expressed as a string of four numbers separated by periods. EC class strings with fewer than four numbers refer to an internal node in the tree, implicitly including all of the subclasses and leaves below it. The numbers specify a path down the hierarchy, with the leftmost number identifying the highest level. All results reported in this paper are based on the EC database ENZYME [4] release 21. Table 1 summarizes the structure of the EC hierarchy. Only those EC classes which have at least one associated protein are used in the current implementation.

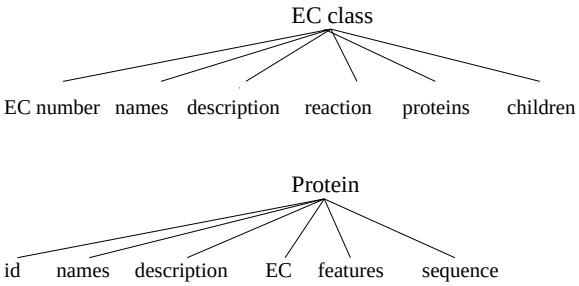

Figure 1: Data representation scheme.

2.1 Data Representation

Each EC class contains a number of text attributes like names, description as well as other complex entities like children, proteins and reaction (figure 1). The EC number uniquely identifies an EC class. To capture the hierarchical nature of the EC a list of child nodes is maintained for each class. Not all attributes are defined over all classes. For instance, the internal nodes of the hierarchy are abstract classes which do not have a specific reaction associated with them. Also, few internal nodes have specific proteins assigned directly to them. On the other hand, leaf nodes always have information about a specific reaction.

In our system, protein data (see figure 1) is parsed from SwissProt records. The ID, name, description and sequence correspond to the fields from SwissProt. Each protein may have one or more EC class numbers assigned to them. These are extracted from SwissProt descriptions.

2.2 Main View

The application is implemented by constructing views which allow intuitive visualization of the data. The main view of the application draws a representation of the EC hierarchy. The textual information for an EC class or protein is shown in separate views. The relationships between the data allow the user to navigate from one view to another.

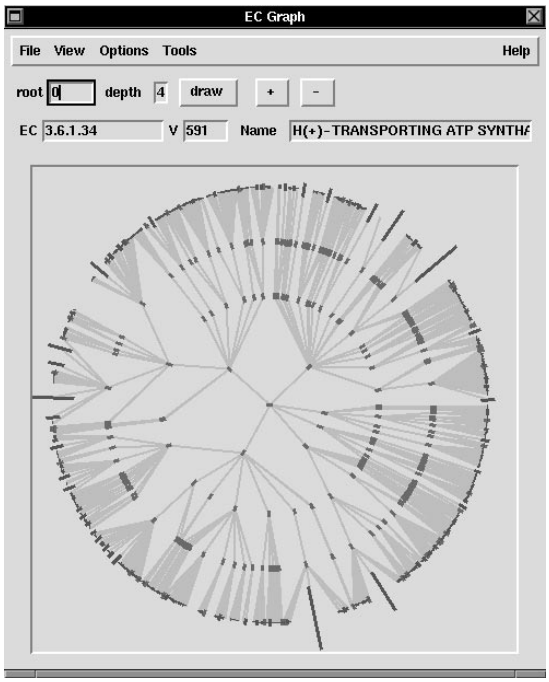

Figure 2: Main view of the application showing the complete EC hierarchy. The length of the nodes is proportional to the number of proteins in each EC class. The node which is currently selected is EC 3.6.1.34, H^+ transporting ATP-synthase, which has 591 proteins.

The core of the application is formed by a visual representation of the EC classification as a tree (figure 2). The leaves of the tree correspond to EC classes while arcs show hierarchical relationships. It is possible to focus on just a sub-tree of the EC by specifying a root EC class in the toolbar (figure 3). Alternatively, the tree can also be magnified (figure 4) to focus into any desired region of interest. The view supports a number of interactions. In the first mode of interaction any given node may be selected directly, retrieving data for the corresponding EC class. The second mode of interaction augments the first one by allowing multiple EC classes to be selected by some similarity criterion. For instance, all EC classes in which NAD^+ is used as a cofactor may be found for further evaluation.

In the last mode of interaction the distribution of quantitative properties across all EC classes can be visualized. This is done by associating the quantitative attributes of EC classes with geometrical properties of the tree[5]. The

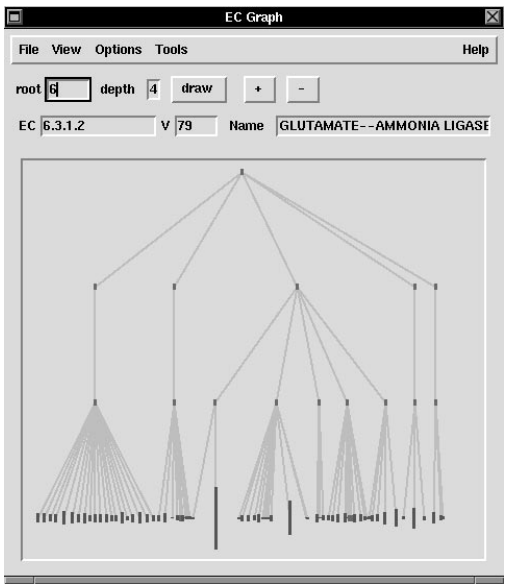

Figure 3: Main view of the application showing the ligases, EC 6. Any sub-tree of the EC can be visualized by entering the root node and the depth in the toolbar. The current selection is EC 6.3.1.2, glutamate ammonia ligase, has the most proteins in all of EC 6.

geometry of the tree is determined by the layout and appearance of the nodes. The layout can be altered by changing the relative position of sibling nodes. Similarly, the color, shape and size of nodes can also be controlled to accentuate quantitative as well as qualitative aspects of EC classes. In figure 2, for instance, each node is shown as lines whose length is proportional to the number of proteins in the EC class.

3 A More Complex Application

We wanted to use this application for visualizing quantitative information about the predictability of each EC class from sequence. We have measured the performance of gapped and ungapped sequence comparison tools, FASTA [6] and BLAST [7], for predicting all 1327 EC classes. The performance was calculated using the ROC statistic [8]. The original data representation scheme (figure 1) was modified to include the sequence alignments for each protein and performance data for each EC class. The degree of correlation between the EC

class and sequences of enzymes is measured by the ROC statistic. When the performance is low it is important to find reasons for the lack of correlation between function and sequence. The schema contains the relevant data for finding these reasons efficiently. Two more views were implemented to visualize the alignments and performance data.

3.1 More Views

The sequence alignment viewer shows a graphical representation of the results of either BLAST or FASTA (figure 6). For each protein match, the SwissProt ID, a schematic of the alignment, similarity score and EC class are shown. Each of the items in this view are active. Clicking on the ID retrieves information about the specific protein. Selecting the aligned regions of a match can be used to retrieve protein data. The aligned regions on the matched proteins may also be selected to view the alignment at the residue level. Finally, selecting the EC class brings the user back to the main view. Besides viewing precomputed alignments, the viewer also allows a new similarity search to be carried out using different parameters or different sequence databases.

The distribution of performance scores is viewed using using the EC visualization tool. To look at ROC curves for EC classes, a separate view is implemented (figure 7). The ROC viewer plots an interactive ROC curve. Selecting a point on the curve retrieves the values of the sensitivity, specificity and similarity score thresholds. Alternatively, the user can specify a particular similarity threshold, sensitivity or specificity, and then see the implications. For example, a user can specify a specificity of 95%, and immediately see the similarity threshold required to generate that specificity for the class, and the sensitivity score that would imply.

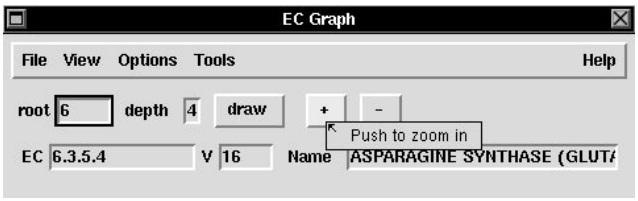

Figure 4: This figure shows the main view toolbar. A small balloon window highlights the button used for zooming into tree. The button labeled '-' is used for zooming out.

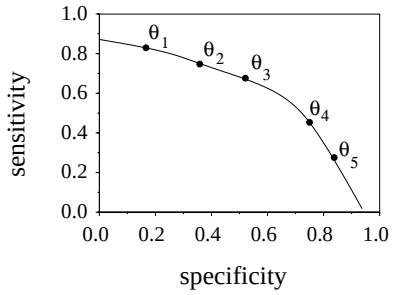

Figure 5: The figure shows an example of a ROC curve. On this figure the sensitivity is shown on the ordinate while the specificity is on the abscissa. Different similarity score thresholds are labeled on the curve as θ_i. Their value increases in the order $\theta_1 < \theta_2 < \theta_3 < \theta_4 < \theta_5$.

3.2 Example

The utility of this application is further demonstrated by an example. Figure 8 shows the EC classification sub-tree under EC 4.2. Enzymes in the sub-class of EC 4.2 are lyases, which catalyze the cleavage of carbon-oxygen bonds. There are three sub-sub-classes under EC 4.2. From the number of nodes in EC 4.2.1, hydro-lyases, it can be seen that they are the largest group under EC 4.2. The performance score for a specific EC class is proportional to the length of the line at each node. The tree also shows the distribution of performance scores for the leaves. Any node in the tree can be selected to view the ROC data. For example, selecting the node for EC 4.2.1.1, carbonate dehydratase, shows that it has a performance score of 0.48. Viewing the ROC curve shows that the maximum sensitivity is around 0.5, even for the smallest similarity score threshold. Hence, there are always false negative matches between proteins in EC 4.2.1.1 at all score thresholds. A quick inspection of the text descriptions of these proteins reveals two evolutionarily distinct forms of carbonate dehydratase. One of these occur in vertebrates while the other are found in plants and prokaryotes. The ROC curve shows that there are non-homologous subgroups in EC 4.2.1.1, which is confirmed by reading the textual information for individual proteins.

Another node in the same tree, dihydroxy-acid dehydratase, EC 4.2.1.9, has a slightly higher performance at 0.67. The ROC curve shows a maximum sensitivity less than 1, which may be characteristic of a class containing non-homologous subgroups. However, inspection of alignments and descriptions of the proteins shows that this is due to short sequence fragments. Because short protein sequences are not specific enough, they have false positive matches

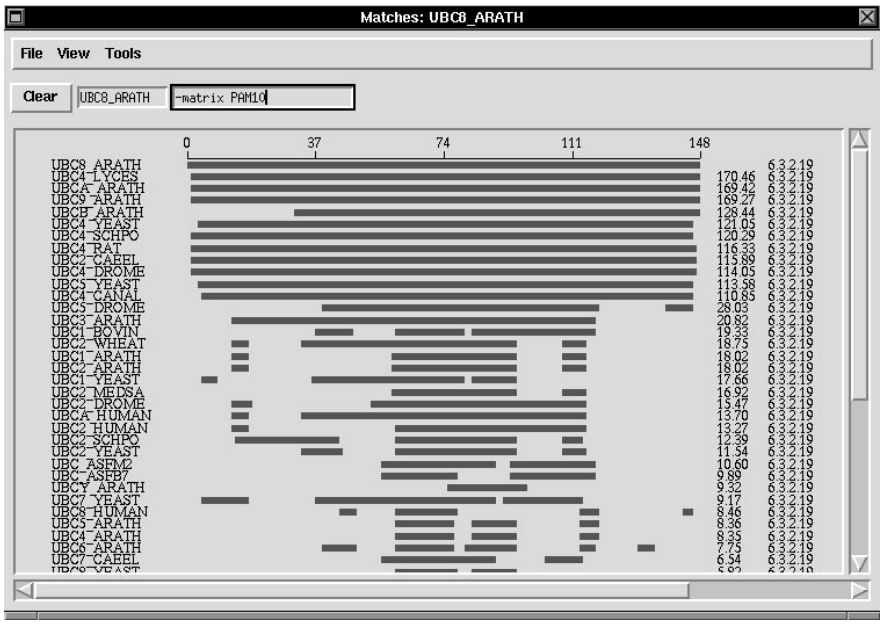

Figure 6: Window showing pair-wise sequence alignments with a protein.

with functionally unrelated proteins.

Two quantitative attributes of EC classes are shown in figure 8. One of these is the performance, which is represented by the length of lines. The second one is the number of proteins available in an EC class, which is rendered as the the amount of horizontal space it takes up in the graph. This geometric quality is computed in the following manner: First, the total number of proteins in a subtree are counted. Second, the fraction of proteins in a sibling node is used to determine the linear space allocated to it. As a result, EC classes with a greater number of proteins are further away from others. For instance, EC 4.2.1.1 has 47 proteins and it occupies more space than its sibling EC 4.2.1.8, for which there only three proteins. In this case, by visual inspection of the relationship between performance and the number of proteins in the class in this area of the hierarchy, it is clear that there is no correlation, implying somewhat counter-intuitively that increasing the number of representatives of an enzymatic family does not make recognizing novel members of that family easier.

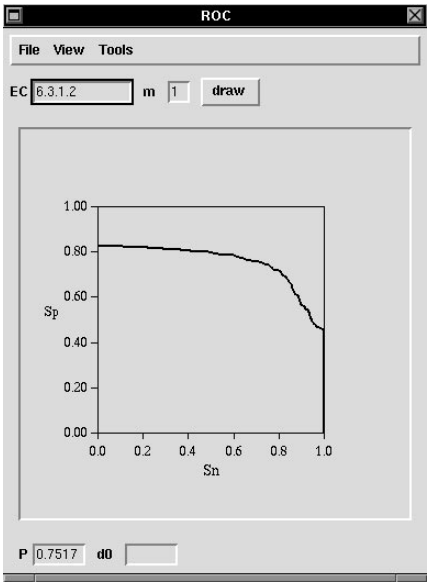

Figure 7: Window showing the ROC curve for an EC class.

4 Conclusion

We developed this tool primarily in order to apply it to visualizing quantitative information about the predictability of each EC class from sequence. The Enzyme Commission nomenclature allows this data to be presented in a hierarchical manner. In this way functionally related enzymes can be presented as sibling nodes in a tree. This tool is useful for visualizing information about large numbers of proteins at once, structured by biochemical activity. Various aspects of these proteins can be mapped onto the activity. The utility of this approach is demonstrated by using it to visualize quantitative data for the performance of protein sequence-based prediction of EC classes. By supporting interactive navigation of data, it helps in finding reasons for the lack of correlation between function and sequence. Our visualization tool enabled us to rapidly identify and explore classes that were difficult to predict for any of a wide variety of reasons. For this application, we extended the protein data structure and added two additional visualizations, one for alignments and one

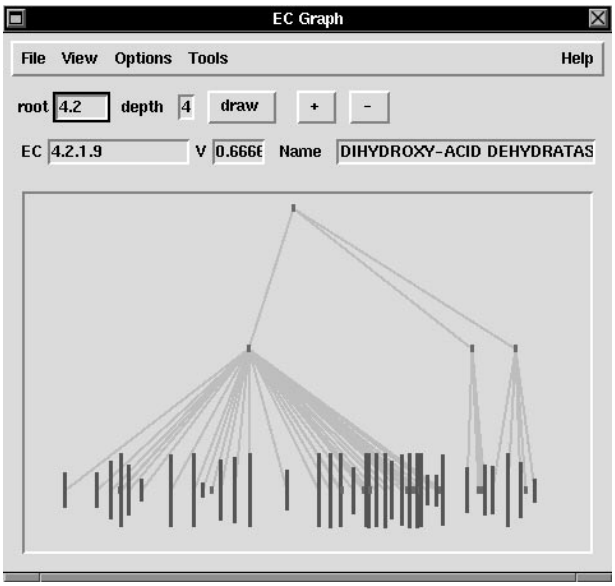

Figure 8: Visualization of subtree for EC 4.2. Focusing on any node automatically describes the EC class by number and name, as well as ROC statistic in the status bar. For instance, the current focus is on EC 4.2.1.9, dihydroxy-acid dehydratase, which has a performance of 0.67. The sub-tree which is being viewed has three levels in which EC 4.2 is the root node. Its children, from left to right, are EC 4.2.1, EC 4.2.2 and EC 4.2.99. The first five children of EC 4.2.1 starting from the left are EC 4.2.1.1, EC 4.2.1.2, EC 4.2.1.3, EC 4.2.1.8 and EC 4.2.1.9. The geometrical properties of a node are associated with two quantitative attributes of an EC class. The length of the line at each node is proportional to the performance; the linear space occupied by a node is proportional to the number of proteins.

for quantitative performance data. The object-oriented design of the visualization tool made these extensions straightforward to implement.

Extensions of this tool may be useful in facilitating other complex analysis of enzymes. Two such problems are genome sequence annotation and metabolic pathway prediction. When putative genes are characterized by sequence alignment, prior knowledge about the reliability of predicting function from sequence can be quite valuable. The performance data we have calculated may be used as such a measure of reliability. With some extensions, this application can be used to for interactively assigning function to genes. A genomic sequence viewer, for instance, would help in visualizing functional annotation. By displaying predicted enzymes on the EC tree as color, and node size for

known performance, the distribution of reliable predictions can be visualized.

Using predicted enzymes, the presence of known metabolic pathways can also be inferred. It can be useful, for instance, to see which steps in a pathway are missing. An additional layer of pathway information can be added to the previous EC tree using a different node type, like circles. In this way, the missing enzymes would be clearly visible over the background of the EC tree, enzymes predicted from a genome and enzymes present in a pathway. Perhaps it might also be possible to generate active visualizations that combined the functionality of this tool with that of metabolic pathway and other reaction databases, perhaps integrating it with a metabolic pathway visualization system.

In conclusion, we believe that the EC hierarchy provides a useful structure upon which visualization can be crafted, and that our tool may be useful to a variety of researchers working in this area.

References

1. I. Shah and L. Hunter. Predicting enzyme function from sequence: A systematic appraisal. *ISMB*, 5:276–283, 1997.
2. Report of the Commission on Enzymes of the International Union of Biochemistry. Pergammon Press,Oxford, 1961.
3. Enzyme Nomenclature. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, Academic Press, New-York, 1992.
4. A. Bairoch. The ENZYME data bank. *Nucleic Acids Res.*, 22:3626–3627, 1994.
5. E.R. Tufte. *Envisioning Information.* Graphics Press, Cheshire, Connecticut, 4 edition, 1994.
6. W.R. Pearson. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, 183:63–98, 1990.
7. S.F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.
8. J.A. Swets. *Measuring the Accuracy of Diagnostic Systems.* Academic Press, New York, 1982.