Pacific Symposium on Biocomputing 6:240-250 (2001)

ViewFeature: Integrated Feature Analysis and Visualization

D.R. Banatao[1, 2], C.C. Huang[1], P.C. Babbitt[1], R.B. Altman[2], T.E. Klein[2]

[1] *Medical Information Sciences and Department of Pharmaceutical Chemistry, University of California, San Francisco 94143-0446, USA*
[2] *Stanford Medical Informatics, 251 Campus Drive MSOB x215, Stanford University, CA 94305-5479, USA*

Visualization interfaces for high performance computing systems pose special problems due to the complexity and volume of data these systems manipulate. In the post-genomic era, scientists must be able to quickly gain insight into structure-function problems, and require flexible computing environments to quickly create interfaces that link the relevant tools. Feature, a program for analyzing protein sites, takes a set of 3-dimensional structures and creates statistical models of sites of structural or functional significance. Until now, Feature has provided no support for visualization, which can make understanding its results difficult. We have developed an extension to the molecular visualization program Chimera that integrates Feature's statistical models and site predictions with 3-dimensional structures viewed in Chimera. We call this extension ViewFeature, and it is designed to help users understand the structural Features that define a site of interest. We applied ViewFeature in an analysis of the enolase superfamily; a functionally distinct class of proteins that share a common fold, the α/β barrel**,** in order to gain a more complete understanding of the conserved physical properties of this superfamily. In particular, we wanted to define the structural determinants that distinguish the enolase superfamily active site scaffold from other α/β barrel superfamilies and particularly from other metal-binding α/β barrel proteins. Through the use of ViewFeature, we have found that the C-terminal domain of the enolase superfamily does not differ at the scaffold level from metal-binding α/β barrels. We are, however, able to differentiate between the metal-binding sites of α/β barrels and those of other metal-binding proteins. We describe the overall architectural Features of enolases in a radius of 10 Angstroms around the active site.

1. Introduction

1.1 Structure-Function Paradigm

As genome projects are completed, the bottleneck of trying to associate structure and function with identified gene targets will remain. The most popular and quickest way to infer function has been through sequence alignments and computer modeling of 3-dimensional structures. Efforts in structural genomics are also addressing this issue by characterizing many new protein folds and structures with techniques such

as x-ray crystallography, NMR, mass spectroscopy, and computer modeling. As protein structures become more abundant, scientists must be able to more quickly and accurately assign function to proteins in fold space.

Detecting conservation in divergent proteins has proven to be a useful way to demonstrate functional evolutionary relationships. Multiple sequence alignments and structural alignments are powerful tools for detecting relationships between proteins, but require a trained eye for detection and interpretation of structural and functional relationships. However, specific differences, across a superfamily of low sequence homology, i.e. the enolase superfamily, are difficult to detect with traditional backbone alignments or sequence alignments.

Huang *et. al*. have previously described a method to infer functional relationships in proteins by linking sequence and structural analysis tools through Chimera, an extensible molecular visualization graphics program[1]. They show that integrated multiple sequence alignment, structural alignment, and visualization tools are applicable in detecting structural relationships in an enzyme superfamily. Another, more direct, approach would be to compare the actual structural elements and biological properties that define the active sites or binding sites versus individual residues.

The Feature program directly compares the biophysical and biochemical elements that make up the 3-dimensional space surrounding sites and non-sites[2]. Feature generates a statistical model from these comparisons and can use this model to predict sites in other structures.

1.2 Structural Scaffolds and The Enolase Superfamily

It is understood that the structural scaffold of some superfamilies is conserved through evolution in order to stabilize a common type of intermediate required by each reaction mechanism[12]. In the enolase superfamily, three acidic residues are conserved in all the members of the superfamily and are associated with a common chemical step important to the function of all of the divergent members[13]. These residues bind a divalent metal cation in the active site that participates in abstraction of an α-proton from carboxylic acids. Describing this scaffold has helped to provide an understanding of how these protein architectures evolved to deliver very different overall chemistries using some common structural and functional elements. The insights obtained are already being applied to re-engineer one member of the superfamily for industrial biocatalysis[14] and show promise for many other applications, including the re-engineering of these enzymes for bioremediation of environmental pollutants or development of new therapies.

The enolase superfamily, also known as the mandelate racemase /muconate lactonizing enzyme /enolase superfamily, is a good model system for creating a structural scaffold. The enolase superfamily is well described in terms of a

superfamily that participates in a diverse array of chemical reactions, substrates, and biochemical functions[3]. Membership in the superfamily is verified experimentally and through structural and sequence alignments. Although, the superfamily is well characterized in many of its various chemistries, there is not yet a detailed description of the structural and functional common elements that make up its overall scaffold. In particular, there remain questions as to which elements can be associated with the partial mechanistic step shared by all members of the superfamily and which elements can be associated with the differences in each member's substrate specificity and overall chemical reactions. Therefore, without a scaffold that includes a description of the conservation observed in the side-chain geometries and interactions with the metal ions that are required for enzymatic activity, it is difficult to quickly assign new members to the superfamily based on a superficial review of its structure.

The Feature program and ViewFeature extension provide the most direct, computational path to gaining physical clues about the enolase superfamily scaffold. We have used Feature to aid in understanding the microenvironments within the C-terminal domain (α/β barrel). The enolase superfamily makes a good candidate for using the Feature system since the well-characterized members will serve as controls for evaluation of the predictive tool, while the less well-characterized members may serve as test cases.

1.3 Feature

Developed by the Helix group at Stanford University, Feature uses a classification or supervised learning algorithm to build statistical models of sites. The algorithm has been described and applied to ion binding sites, enzymatic active sites, and small molecule binding sites[2, 4]. Sites are regions within a protein defined by a central location and a surrounding neighborhood. More specifically, an atom within the region of inquiry can be chosen as the center of a neighborhood with a user-specified radius. Sites are usually picked because of their structural or functional role, such as enzymatic active sites or Ca++ binding. Thus, a nonsite may be described as any other random site where a different function or lack of function may take place. For example, when testing Ca++ binding sites, Mg++ binding sites may be used as nonsites. Feature then takes the defined site region, computes the spatial distributions of biophysical and biochemical properties, and then reports those regions within a site where these properties significantly vary from those of control nonsites. Feature uses a non-parametric test (Mann-Whitney rank-sum test) to find for which properties at which respective volumes are the known positive sites significantly different from the negative control sites. Distinguishing properties are plotted in a 2-D array. The properties implemented include for example: atom type, chemical groups, amino acids, secondary structures, charge, polarity, mobility, and

solvent accessibility. Feature then uses a log-odds scoring function based on Bayes' Rule to obtain the distribution of distinguishing properties in a query structure[4]. Feature gives a score that indicates how likely a query region is a site of interest. We are currently using the Unix version of the Feature code written in C++.

1.4 ViewFeature and Chimera Extensibility

Feature typically displays distinguishing properties in 2-D plots and marks potential sites in the Kinemage[6] format. By integrating Feature analysis into Chimera, we are enabling visualization of those distinguishing properties in the actual 3-D structure within the specified volumes. ViewFeature highlights significant properties, thus giving users insight into the structural motifs that define a given active site or protein structural scaffold.

We wrote ViewFeature as an extension to Chimera, a molecular visualization graphics program developed by the UCSF Computer Graphics Laboratory[5]. Chimera is written in C++ and the Python programming languages to provide extensibility to users and uses the Object Technology Framework object class library for manipulating molecular data. The Python programming language enables our extension modules to be platform independent. Chimera also uses Tk and OpenGL for its graphical user interface and 3-dimensional graphics. The most recent releases of Chimera currently run on Unix and PC platforms, with Linux versions in development.

ViewFeature displays statistical data for distinguishing properties, determined by Feature, in an interactive viewer. This viewer allows the user to manipulate a protein model and highlight all atoms that relate to distinguishing properties. To better orient the user in 3-dimensional space, ViewFeature provides visualization of the specific spherical volume for a volume/property pair oriented around a specific site. ViewFeature, when applied to predicted sites in a query structure, shows which exact properties contributed to Feature building up a site at that location. This visualization technique facilitates a user's perception of a protein's fold, active sites, and general architecture. With multiple protein models open, ViewFeature provides a way to visualize the direct comparison between sites in different models. In the ViewFeature interactive viewer, positive training sites and predicted sites can be selected for the respective models. Choosing a property/volume pair in the viewer will highlight that pair in the open models, respective to the site point. This visualization of Feature's statistical output provides a new way to distinguish protein sites from one another.

2. Methods

2.1 Feature Analysis of the Enolase Superfamily

We began by choosing high-resolution crystal structures (resolution greater than or equal to 2 Angstroms) containing sites of interest from the Protein Data Bank (http://www.rcsb.org/pdb)[8]. We chose representative structures for mandelate racemase, muconate lactonizing enzyme, and enolase for the positive training set. The use of a control group as the baseline for statistical testing is a critical element of the Feature method. For example, we tested the hypothesis that the enolase superfamily was significantly different from other α/β barrels. Thus we chose a representative set of non-homologous, metal-binding α/β barrel proteins as well as non-metal-binding α/β barrel proteins to serve as the control training. Proteins were checked for sequence and structure homology using the PDBSelect[10, 11] and SCOP databases[9].

Sites were specified as three-dimensional locations at the center of a 10-Angstrom radius sphere. Positive sites were specified as the x-y-z coordinates of the metal ion, since these atoms are essential to the activity of all enolase superfamily members and is the most easily distinguished common element for all members. From the control set of proteins, negative examples of sites, or nonsites, could be picked explicitly or through random sampling based on surrounding all-atom density. We specified nonsites where a lack of function was likely at that location. For example, when comparing enolase superfamily proteins to metal-binding α/β barrel proteins, we chose the location of the Mn++ ion as a positive site. Negative sites were other metal ions or Mn++ ions in a non-enolase superfamily protein. To pick random sites based on density, we first calculated the average all-atom density, within a site radius, of all positive training sites. We then scanned the control training proteins and created a list of coordinates for each protein, where the surrounding all-atom density was within one standard deviation of the average density of the positive sites.

We ran Feature's training algorithm on each training set. Feature produced a list of distinguishing properties that described a statistical model for each set. To evaluate the specificity of these statistical models, we used Feature's scanning algorithm to predict sites in query structures from a scanning set (those proteins left out of the training set). This set was divided in two groups. The control scanning set contained a consistent set of proteins and was scanned after each training experiment. This set consisted of enolase members, metal-binding α/β barrel proteins, non-metal-binding α/β barrel proteins, and representative structures from each of the major classes of the SCOP database. The other group of scanning

proteins varied with each experiment and served as positive and negative controls for each experiment. This group consisted of positive training proteins and control training proteins from the respective training set.

POSITIVE SITES			
Protein Name	**PDB #**	**# Sites**	**METAL**
muconate lactonizing enzyme	1muc	3	MN
enolase	1one	3	MG
mandelate racemase	2mnr	3	MN
	TOTAL	**9**	
NEGATIVE SITES			
beta-mannanase	1bqc	20	
cyclodextrin glycosyltransferase	1cgt	22	CA
seed storage protein	1cnv	20	
fructose-bisphosphate aldolase	1dos	19	ZN
endoglucanase CelA	1edg	20	
alpha amylase	1hny	19	CA
indole-3-glycerophosphate	1pii	20	
tRNA-guanine transglycosylase	1pud	19	ZN
bacterial chitobiase, catalytic domain	1qba	20	
xylanase A, catalytic core	1tax	20	
d-xylose isomerase	1xis	21	MN
d-xylose isomerase	1xya	17	MG
endo-beta-N-acetylglucosaminindase	2ebn	17	ZN
ribulose 1,5 bisphosphate carboxylase	3rub	20	
	TOTAL	**274**	

Table 1. Training Set of a Feature experiment comparing Enolase Superfamily members, 1muc, 1one, and 2mnr against metal and non-metal-binding α/β barrel proteins. γ or δ Carbon atoms of metal-binding residues in the active site were used as positive sites, while non-sites were chosen based on surrounding all-atom density within a given radius, or proximity to a metal atom. Some of the proteins from this training set were later used as controls for scanning.

2.2 ViewFeature Implementation and Analysis

Chimera's extensibility allows ViewFeature to access full control of Chimera commands without having to modify any Chimera source code. We were also able to take advantage of Chimera's handling of molecules and data. This enabled us to write scripts that could fully communicate with Chimera to display and visualize the properties defined in Feature. The graphical user interface (GUI) for ViewFeature was written in Tkinter, for which Python has an interface. This allowed our GUI to have custom menus and dialog boxes, which we have built to aid users in the file

handling of Feature output files. Easy to use and platform independent file management for Feature will be a requirement as Feature and Chimera move to various platforms.

After Feature scanned through query proteins, it provided scores to indicate how likely a position is an actual site. We compared scores between positive and negative control query proteins. The deviation in scores between true positives and true negatives was an indication of how well a training set was put together. If the deviation was greater than 20, we examined the training set further in ViewFeature.

PDB models were imported into Chimera. We then used ViewFeature to open all relevant Feature data files pertaining to the open protein model. ViewFeature mapped distinguishing properties to specified volumes and sites. Integrated Feature data with protein structures allowed us to understand the collective effect properties had in distinguishing one set of proteins from another. In these experiments, for example, we were able to see what differentiated the metal-binding α/β barrels from other metal-binding protein families.

3. Results

Our Feature analysis of the enolase superfamily did not reveal significant differences between the enolase superfamily and other metal-binding α/β barrels. However, there was a detectable difference between metal-binding α/β barrels (including enolases) and other divalent metal-binding proteins.

Our analysis showed that the inner barrel of enolase superfamily proteins and other metal-binding α/β barrel proteins is a charged and acidic environment. Within 10 Angstroms of the metal ions bound inside the active site of metal-binding α/β barrels are many solvent accessible residues. There is a prevalence of glutamate residues within 4 Angstroms of the metal ion. Other highly charged residues stem from the beta sheets lining the α/β barrel and point inwards toward the metal ion. There is a lack of non-polar residues at the center of the barrel. Helical secondary structure is deficient near the active site's beta-barrel region. Beta sheets begin at 5 Angstroms from the metal ion and extend past 10 Angstroms away. These findings are consistent with what is currently known about the enolase superfamily.

Our findings suggest that the enolase superfamily active site scaffold is similar to other divalent-metal binding α/β barrel proteins. Our preliminary results also suggest that α/β barrel proteins bind metals differently than other divalent metal-binding proteins.

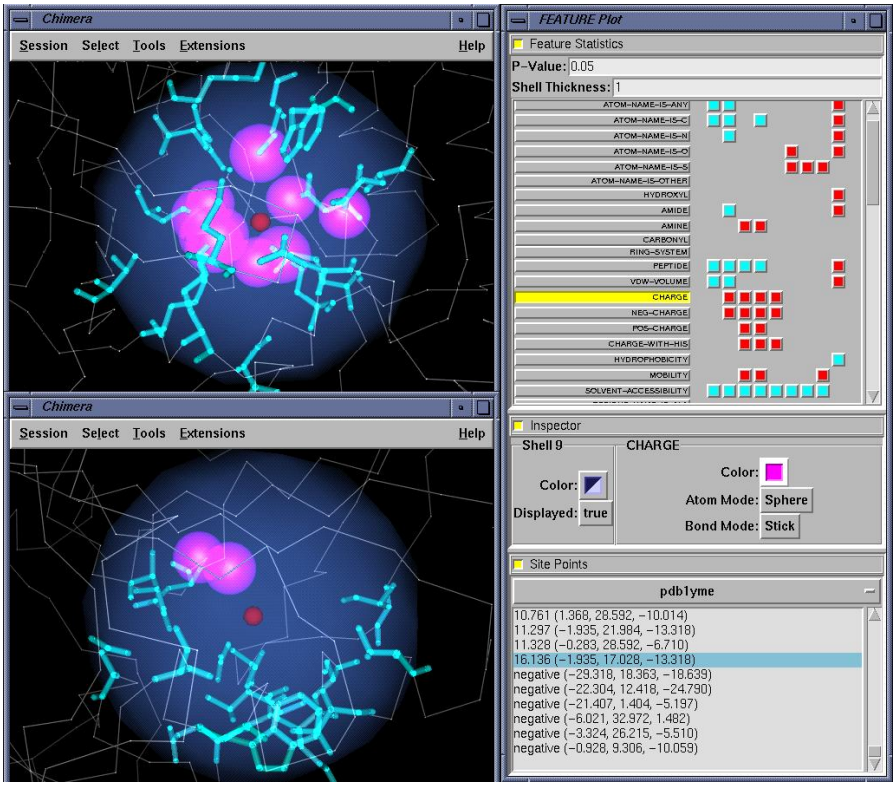
4. Conclusion

We have written an extension to Chimera for integration and visualization of Feature statistical data, site predictions, and 3-dimensional structures. As secondary and tertiary structure information about proteins serves as input to Feature, the usefulness of being able to visualize Feature's output mapped onto the associated 3-dimensional structure became quickly apparent.

As protein structures and models become abundant, scientists need efficient and objective methods to analyze structures and assign function. Feature is a method that allows analysis of multiple structures in an automated fashion. ViewFeature translates Feature's statistical results into an intuitive, interactive visual media. ViewFeature also benefits from the ability to incorporate all the functionality of the Chimera visualization tool when performing structural analysis. Chimera is highly extensible and promotes integration of structure and sequence analysis tools. The integration of these different techniques creates a powerful method for future protein structure analyses.

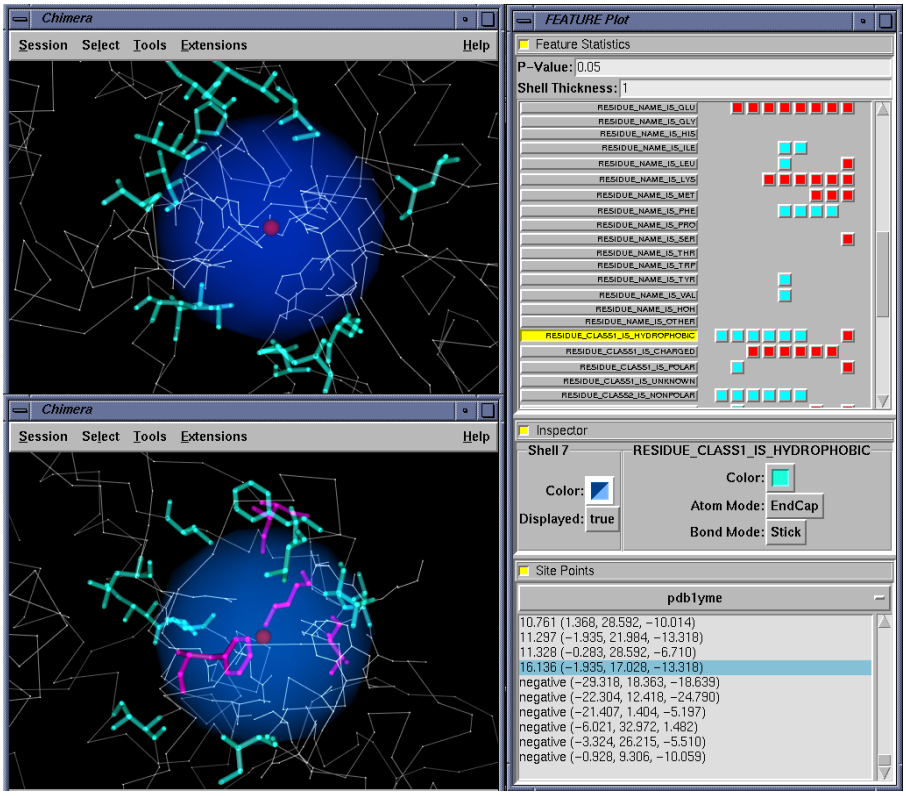
In order to demonstrate the potential value of ViewFeature we have used it in a preliminary analysis of the enolase superfamily, in an attempt to understand the Features that distinguish this superfamily from superfamilies of α/β barrels. Using the currently available structures for the enolase superfamily, Feature did not detect major differences between the active site scaffolds of enolase superfamily proteins and other metal-binding α/β barrels. The analysis of this superfamily may require more example structures to provide statistical power for detecting subtle differences. However, Feature was able to detect differences between the sites of metal-binding α/β barrels and other metal-binding proteins. Wei and Altman have previously shown the ability to predict Ca++ binding sites in proteins using Feature[4]. We have shown that Feature can go further to discriminate between divalent metal-ion binding sites of α/β barrels and other divalent metal-binding proteins. This demonstrates an ability of Feature to detect properties of interest across fold classifications.

In addition to visualization, ViewFeature was useful in organizing and handling the large amount of statistical data and output files of Feature. More and more, biology is moving toward being a high-throughput, quantitative science. New methods for organizing and manipulating the flow of scientific information will continue to be increasingly necessary. These techniques will enable scientists to be more efficient whether working at the bench or the computer.



Results of Feature analysis

ViewFeature displays Feature's finding that the active site region of enolases (and metal-binding TIM barrels) are statistically significant from other divalent metal-binding proteins. The Chimera viewers (left panels) show enolase superfamily member, chloromuconate cycloisomerase (2chr, top), and Zn-dependent exopeptidase, carboxypeptidase A (1yme, bottom). Right panel shows ViewFeature's interactive viewer. Statistical data from Feature is plotted in the Feature Statistics panel (upper right). Properties in red are statistically significant in the positive training set. Properties in cyan are statistically deficient in the positive training set (significant in the negative training set). Shells and properties can be manipulated in the Inspector panel (middle right). Training and scanned sites are managed and displayed through the Site Points panel (lower right). In this experiment, we trained Feature to detect enolase superfamily active sites against metal-binding proteins. 2chr was excluded from the training set. The highest scoring predicted sites for 2chr and 1yme are displayed as red spheres. Solvent accessible residues were highlighted in cyan by selecting the corresponding property in the statistics panel. Charged atoms were highlighted as large, magenta spheres. The statistics panel shows that charged atoms are significant 2-6 Angstroms from metal ions in enolase superfamily members.



Differences in metal-binding sites

The picture highlights another statistical difference between the metal-binding sites of TIM barrels and other divalent metal-binding proteins. Hydrophobic residues are highlighted by clicking, “RESIDUE_CLASS_IS_HYDROPHOBIC”, in the statistics panel. In this diagram, the shell at 7 Angstroms is shown as a transparent, blue sphere. Hydrophobic residues closer than 7 Angstroms are colored magenta. All other hydrophobic residues between 7-10 Angstroms are colored cyan. The highest scoring site predicted by Feature is shown as a small, red sphere.

Acknowledgements

The authors gratefully acknowledge the UCSF Computer Graphics Laboratory (Dr. Tom Ferrin, P.I.) as well as financial support from the NIH National Center for Research Resources (P41 RR01081), DOE DE-FG03-99ER62269, NIH RO1 grant GM60595, NIGMS Biomedical Science Research Career Enhancement Program, and the UCOP training grant.

References

1. C.C. Huang, W.R. Novak, P.C. Babbitt, A.I. Jewett, T.E. Ferrin, T.E. Klein, “Integrated Tools for Structural and Sequence Alignment and Analysis”, *Pacific Symposium on Biocomputing* 230, (2000)
2. S.C. Bagley, R.B. Altman, “Characterizing the Microenvironment Surrounding Protein Sites”, *Protein Science*, 4, 622 (1995)
3. P.C. Babbitt, G.T. Mrachko, M.S. Hasson, G. W. Huisman, R. Kolter, D. Ringe, G.A. Petsko, G.L. Kenyon, J.A. Gerlt, “A Functionally Diverse Enzyme Superfamily That Abstracts the a Protons of Carboxylic Acids”, *Science* 267, 1159 (1995)
4. L. Wei, R.B. Altman, “Recognizing protein binding sites using statistical descriptions of their 3D environments”, *Pacific Symposium on Biocomputing* 497, (1998)
5. C.C. Huang, G.S. Couch, E.F. Pettersen and T.E.Ferrin, "Chimera: An Extensible Molecular Modeling Application Constructured using Standard Components", *Pacific Symposium on Biocomputing* 724, (1996)
6. D.C. Richardson, J.S. Richardson, “The kinemage: A Tool For Scientific Communication”, *Protein Science* 1, 3 (1992)
7. G.A. Petsko, G.L. Kenyon, J.A. Gerlt, D Ringe, J.W. Kozarich, “On the Origin of Enzymatic Species”, *Trends Biochem Sci*. 10, 372 (1993)
8. F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.F. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, “The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures”, *Arch Biochem Biophys.* 185, 584 (1978)
9. A.G. Murzin, S.E. Brenner, T. Hubbard, C.Chothia, “SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures”, *J. Mol. Biol.* 247, 536 (1995)
10. U. Hobohm, M. Scharf, R. Schneider, C. Sander, “Selection of a Representative Set of Structures from the Brookhaven Protein Data Bank”, *Protein Science* 1, 409 (1992)
11. U. Hobohm and C. Sander, “Enlarged Representative Set of Protein Structures”, *Protein Science* 3, 522 (1994)
12. P.C. Babbitt and J.A. Gerlt, “Understanding Enzyme Superfamilies: Chemistry as the Fundamental Determinant in the Evolution of New Catalytic Activities”, *J. Biol. Chem.* 272, 30591 (1997)
13. P.C. Babbitt, M. Hasson, J.E.Wedekind, D.J. Palmer, M.A. Lies, G.H. Reed, I. Rayment, D. Ringe, G.L. Kenyon, J.A. Gerlt, “The Enolase Superfamily: A General Strategy for Enzyme-Catalyzed Abstraction of the α-protons of Carboxylic Acids” , *Biochem.* 35, 16489 (1996)
14. A. Bommarius, personal communication