STATISTICAL AND VISUAL MORPH MOVIE ANALYSIS OF CRYSTALLO-GRAPHIC MUTANT SELECTION BIAS IN PROTEIN MUTATION RESOURCE DATA

WERNER G. KREBS and PHILIP E. BOURNE[1,*]

[1]Department of Pharmacology
Integrative Biosciences, San Diego Supercomputer Center Dept 0505, University of California, San Diego, 9500 Gilman Drive
La Jolla, CA 92093-0505, USA

wkrebs@sdsc.edu, bourne@sdsc.edu*

*To whom correspondence should be addressed.

Structural studies of the effects of non-silent mutations on protein conformational change are an important key in deciphering the language that relates protein amino acid primary structure to tertiary structure. Elsewhere, we presented the Protein Mutant Resource (PMR) database, a set of online tools that systematically identified groups of related mutant structures in the Protein DataBank (PDB), accurately inferred mutant classifications in the Gene Ontology using an innovative, statistically rigorous data-mining algorithm with more general applicability, and illustrated the relationship of these mutant structures via an intuitive user interface. Here, we perform a comprehensive statistical analysis of the effect of PMR mutations on protein tertiary structure. We find that, although the PMR does contain spectacular examples of conformational change, in general there is a counter-intuitive inverse relationship between conformational change (measured as C-alpha displacement or RMS of the core structure) and the number of mutations in a structure. That is, point mutations by structural biologists present in the PDB contrast naturally evolved mutations. We compare the frequency of mutations in the PMR/PDB datasets against the accepted PAM250 natural amino acid mutation frequency to confirm these observations. We generated morph movies from PMR structure pairs using technology previously developed for the Macromolecular Motions Database (http://molmovdb.org), allowing bioinformaticians, geneticists, protein engineers, and rational drug designers to analyze visually the mechanisms of protein conformational change and distinguish between conformational change due to motions (e.g., ligand binding) and mutations. The PMR morph movies and statistics can be freely viewed from the PMR website, http://pmr.sdsc.edu.

1. Introduction

Rational drug design seeks to use knowledge of a protein's three-dimensional chemical structure as a target against which to design new drugs. This particular application of x-ray crystallography has likely been one of the principal economic factors driving growth in the experimental field. However, mutant proteins may occur naturally in the target host population or, in the case of drugs such as antibiotics, may evolve in the parasite as a

form of drug resistance. Scientists would like to have an understanding of how a putative drug interacts not only with the wild-type protein, but also with its likely mutants. The cost of experimental determination of mutant structures is often still prohibitive. An extensive structural database of proteins and neighboring mutants can be expected to assist scientists in visualizing likely structural changes brought about by mutation and would likely find immediate application in rational drug design through improved homology modeling, which is of importance in the pharmaceutical industry [1-4].

The deduction of a detailed, three-dimensional chemical structure of a protein from its genetic sequence is a fundamental and long-studied problem in structural biology [5-7], of importance in *de novo* structure prediction, protein folding, crystallographic refinement phasing, molecular dynamics, and computational chemistry. At present, it is solved principally through the labor-intensive but effective process of X-ray crystallography and NMR. A direct study of existing data on the effects of protein mutation on protein structure can have immediate payoffs [7, 8].

Although databases of mutant gene products [9-12] as well as specialized databases of mutant protein structures have previously been developed [13-15], the PMR [16] was the first PDB-wide [17] database of mutant protein structures. Entering the PDB ID of a structure in the PMR into the entry form on the PMR home page brought up the sequence of the wild-type structure for that mutant family along with a listing of the differences in amino acid sequence between the wild-type and the selected PDB ID (Figure 1). Users could click on any of the mutation sites listed for the wild-type structure to obtain a listing of the available mutant structures with modifications at the amino acid position (Figure 1). An anticipated use of the PMR was in protein engineering. Scientists seeking to modify the sequence of an existing protein to express a slightly different structure could search the PMR for sequences matching the protein's current sequence and then examine the stored mutations for structure variants [18, 19].

The PMR website had a number of other innovations; in particular, the PMR GO classification feature utilized an improved means of and database-wide statistically rigorous gene annotation and data-mining with widespread applicability; it employed a new method for statistical inference that reduced false positives by more than 10% over similar existing methods, which we have presented elsewhere [20, 21]. This comprised an innovative statistical algorithm for mapping new entities (genes) in a database into an ontological classification when the database can be clustered in a way related to the ontological classification and a subset of the database has previously been mapped into the ontology. The method could be used to enhance annotations derived from complementary automatic methlologies, including as text-mining [22-25]. At the time we were able to classify approximately $2/3^{rds}$ of the PDB using this method. Since then, we have found that iterative application of our method, in conjunction with additional starting data, allows the method to classify nearly the entire PDB. We have developed a new, more efficient SQL implementation of our $O(N \lg N)$ algorithm to allow streamlined execution within a scalable database engine. Originally developed in Perl using an Oracle database as the backend datastore [20], we desired a more efficient, scalable implementation to support application of the algorithm to larger databases, such as Genbank. This involved

re-implementation the algorithm largely as a sequence of SQL commands performing internal counting, indexing, and sorting operations inside an IBM DB2 database instance to compute the parameters from an input dataset. We implemented the actual floating-point statistical calculations in an efficient Perl implementation, which requires querying the database for these parameters after they have been computed by the database engine. In principle, however, these parameters can also be computed in SQL using stored procedure calls, allowing the entire algorithm to be run inside the database engine. This should enable a highly scalable implementation by means of a parallel database engine, such as Parallel DB2. Assuming the size of the ontology remains constant, worst-case running time with disjoint clusters is O(N lg N) in terms of the size of the input dataset, so the major computational obstacle likely lies in obtaining the initial clusters.

PMR database entries interacted with a number of external databases (MolMovDB [26-29], GO [30], PubMed/Entrez [31], PDBsum [32]) as well as the PDB. Consequently, the PMR could be used as a portal by those studying families of proteins of closely related sequence within the PDB.
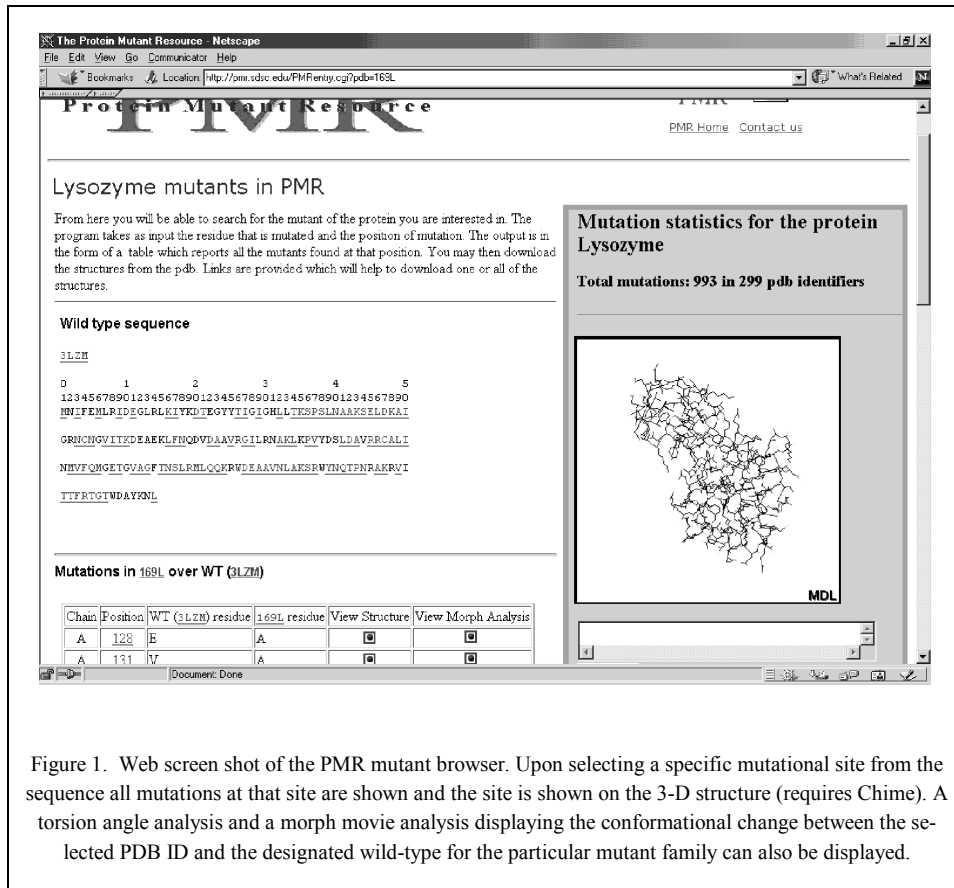
Figure 1. Web screen shot of the PMR mutant browser. Upon selecting a specific mutational site from the sequence all mutations at that site are shown and the site is shown on the 3-D structure (requires Chime). A torsion angle analysis and a morph movie analysis displaying the conformational change between the selected PDB ID and the designated wild-type for the particular mutant family can also be displayed.

Here, we characterized the effect of PMR mutations on protein tertiary structure statistically and detected a potential selective bias in available PMR/PDB structures of mutant proteins. To confirm this, we compared the frequency of mutations in PMR/PDB datasets against accepted PAM 250 natural amino acid mutation frequencies. We further improved on the PMR web interface by generating and making morph movies of the conformation changes available on the web. These automatically generated morph movies assist scientists in visually discriminating between conformational changes caused by protein motions [29] from those caused primarily by changes in protein sequence [16]. We believe both our statistical analysis of PMR/PDB data and our morph movies of PMR data will be of general interest to the structural bioinformatics community. Our morph movies are freely available off the PMR website (http://pmr.sdsc.edu).

Distribution of Mutations

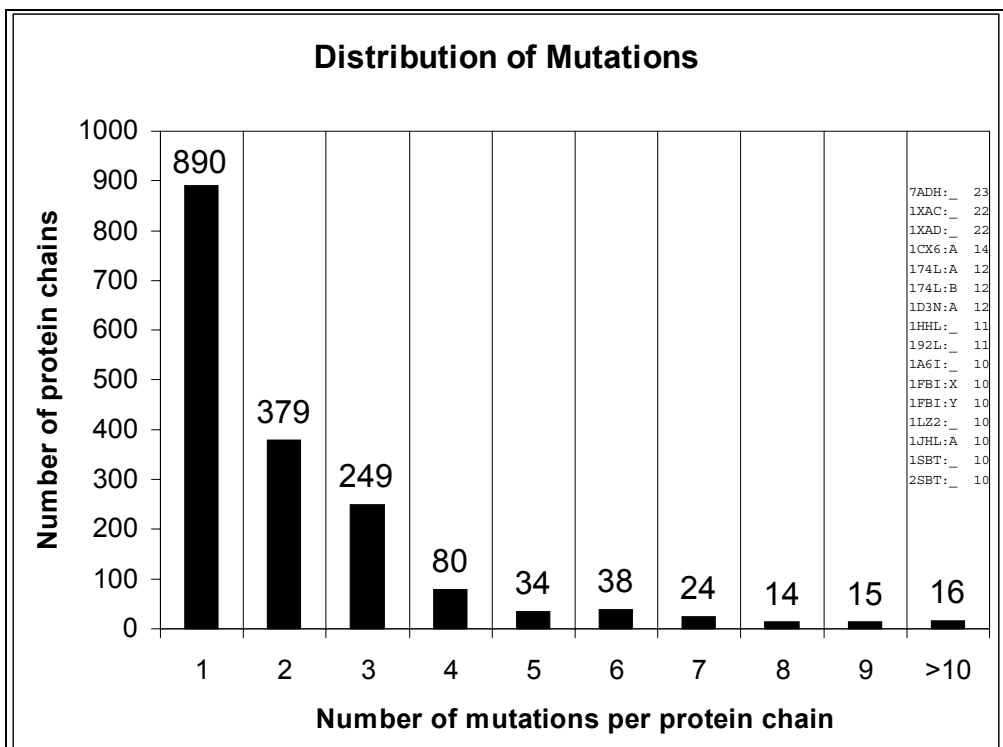|	
7ADH:_	23
1XAC:_	22
1XAD:_	22
1CX6:A	14
174L:A	12
174L:B	12
1D3N:A	12
1HHL:_	11
192L:_	11
1A6I:_	10
1FBI:X	10
1FBI:Y	10
1LZ2:_	10
1JHL:A	10
1SBT:_	10
2SBT:_	10

Figure 2. Distribution of the number of mutations per polypeptide chain in the PMR. The specific PDB and chain identifiers for each polypeptide chain with > 10 mutations are shown.

Figure 3. Mean Cα displacement by mutant residue type for polypeptide chains with single mutations.
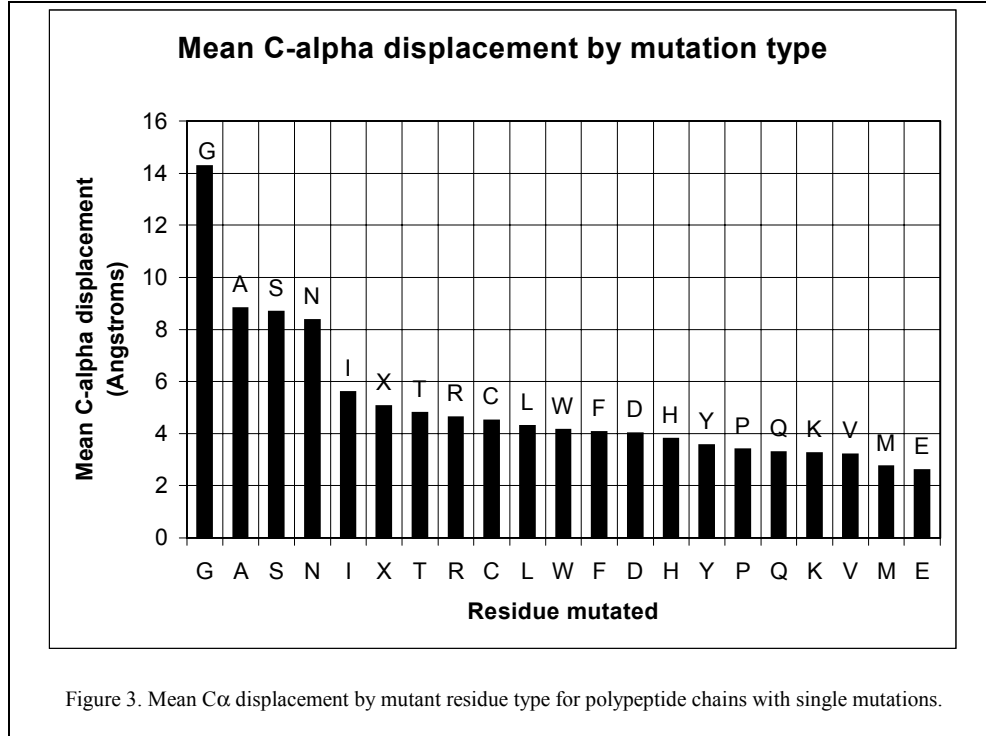
2. Materials and Methods

As described elsewhere [16], the PMR was generated by automatically clustering the PDB [17] at 95% sequence identity using the CD-HIT sequence clustering approach [33]. CD-HIT uses a greedy algorithm [34] to sort and process sequences in order of decreasing length; the longest sequence in each cluster becomes its representative. Efficiency is achieved because sequences are processed by comparing them only against the representative sequences for each established cluster to decide whether they should be added to an existing cluster or become the representative for a new cluster.

The resulting clusters were then manually filtered into species-based families and a 'wild-type' PDB chain was manually selected from each family by inspection of the scientific literature [16]. Software was developed to automatically find and add new PDB entries to existing PMR families on a regular basis. The resulting data were loaded into Oracle tables and originally made freely accessible via a web interface, originally developed in Perl [35], Oracle SQL, Chime, JavaScript, and HTML.

Table 1. Mutation statistics for the PMR database taken from 1157 PDB structures.	
Total number of mutations (chains)	3343
Wild-type PDB chains	194
Non wild-type PDB chains	3149
Mutation Sites	1157
Average number of residues mutated per chain	2.9
Most commonly mutated amino acid in PMR	Alanine

Morph movie technology [26, 27] originally developed for the Database of Macromolecular Motions [26, 28, 29, 36, 37] (http://molmovdb.org) was applied to PMR data to make four-dimensional visual illustrations [26, 27, 38] of the conformation changes induced by mutation. Due to the human eye's evolutionarily-induced sensitivity to moving objects, morph movies are useful way of visualizing the difference between similar protein structures. Structural change is given in terms of $C\alpha$ displacement for each possible, non-redundant wild-type and mutant structure pair in each PMR family using a sieve-fitting superposition technique [39] which interpolates between solved structures to provide a visual rendering. Protein motions are available as animated GIF images on the PMR website (http://pmr.sdsc.edu; Figure 1). Summary statistics on the types of mutation and the motions induced are given in Table 1. The number of mutations per structure is given in Figure 2. $C\alpha$ displacements by mutant residue type in PMR data are given in Figure 3. The frequency of amino acid mutations in PMR data were tabulated and compared with the accepted PAM250 natural amino acid mutation frequencies [40] and are shown in Table 2 and Figure 4. A scatter-plot depicting the relationship of number of mutations with structural change (as measured by previously computed $C\alpha$ displacements) is shown in Figure 5. A Java web applet for visualizing mutation in terms of color-coded changes in inter-residue carbon contacts is shown in Figure 6.

The web server has recently been ported from Perl to Java 2 Enterprise Edition (J2EE) to take advantage of the greater speed, scalability, platform independence, robustness, and maintainability of a three-tiered J2EE web server design. The three-tiered, industry-standard struts framework design separates code dedicated to content presentation, database back-end interaction, and website navigation into separate, object-oriented modules, allowing simultaneous, independent development and maintenance of these different aspects of the web site by programmers with different skills. The resulting web server code is also much more scalable, because code dedicated to database interaction ("middle-tier code") runs in a different set of threads than code dedicated to content presentation, allowing these functions to run on separate servers. Unlike a convention two-tiered web design, where each web browser effectively requires the attention of a database listener thread, a single middle-tier thread requires at most a single database listener

thread, yet can handle many simultaneous presentation tier threads and their associated web browsers. Scalability is further achieved through pooling of database listener threads and their associated database connections among middle-tier threads. The J2EE server version of the PMR now obtains sequence, space group, and structure coordinates from the PDB by means of remote procedure calls using the industry standard XML and HTTP-based SOAP web services protocol (Figure 7).

3. Results

A morph analysis of all possible structure pair combinations within PMR families would have $O(N^2)$ complexity with the size of family in terms of both disk space and CPU time and was not computationally tractable even at a supercomputer center. Instead, we limited our analysis to non-redundant combinations of the members of each family and its wild-type. This has $O(N)$ complexity and reduced the task by several orders of magnitude. Running in parallel, our morph server software required three days of CPU time on a sixteen-CPU cluster of four-CPU Sun Ultra-80 servers running SunOS 5.7. The resulting morph movies and statistical data require 4.3 gigabytes of disk storage.

The distribution of numbers of mutations per polypeptide chain is given in Figure 2. Greater than 99% of the PMR mutant structures have 9 or fewer mutations when compared to the wild-type. The rapid fall-off in number of mutations is expected as these structures are usually studied to understand the impact of single or a small number of correlated mutations. Alanine is the residue most commonly used to mutate structures (Table 1) presumably to change the functional role of a given residue and through the presence of a C-beta carbon still confer side chain directionality and some sense of side chain volume in a neutral substitution. Single mutations involving glycine result in the largest structural changes (Figure 3) presumably as a result of significant stereochemical change in side chain volume and possibly physicochemical properties. In theirs of their propensity to cause structural changes in mutations involving them, the amino acids can be separated into at least three clear groups —glycine, a second group involving alanine, serine, and asparagine, and the remaining amino acids (Figure 3).

There is a significant difference in the mutational frequencies for PMR versus the natural PAM250 frequencies (Figure 4 and Table 2). For example, the table shows that the mutation of phenylalanine to tyrosine seems disfavored by structural biologists despite a relatively high frequency during the course of evolution. Conversely, the conversion of alanine to cysteine, uncommon during the course of evolution, is favored by structural biologists, presumably to add stability to a structure under study. Comparison of Figure 3 with Figure 4 and Table 3 suggests that some of the amino acids most commonly involved in structural biological mutation experiments (e.g., alanine, asparagines, serine) are generally associated with mutations resulting in the largest structural changes. Divergence from natural frequencies of mutation is not surprising since structural biology is often concerned with the engineering of proteins to test function. Not surprisingly, loss of function is less common in nature than it is in the laboratory. It is not difficult to engi-

neer a protein that retains its structure, is biologically inactive or has a significantly reduced activity.

Given the complete body of structural data showing mutation from the wildtype, contrary to what one would expect, structural change shows an inverse relationship to the number of mutations (Figure 5) – large structural changes are induced by small numbers of mutations, whereas a large number of mutations can lead to relatively small changes. Consider several specific cases shown in the PMR data (Figure 5). The largest structural change in the PMR occurs in TAQ DNA Polymerase [41]. A number of structures of TAQ DNA polymerase have been solved with an alanine instead of the wild-type glycine at position 152. These all show an unusually large 140Å Cα displacement. Crystal contacts as a cause may be ruled out since the mutant structure has been solved in a number of different space groups, including the same space group as the wild-type, and all structures give an identical Cα displacement. The 140 Å Cα displacement structures are all bound to DNA, whereas the 1TAQ wild-type [41] is an unbound structure indicating that the unusually large conformation change here is likely due to the combined effect of a DNA clamping protein motion as well as the single-point mutation. Thus the PMR does not distinguish between conformational changes induced by ligand binding or complex formation and that induced simply by point mutation. However, large changes induced by point mutations alone do occur. Analysis of the structural change between 7ADH [42] and the designated wild-type alcohol dehydrogenase structure 1ADG [43] shows a displacement of 66Å caused by structural changes resulting from 22 mutations. Visualization of the morphing shows chain breakages and backbone elements passing through each other. That is, unlike TAQ DNA polymerase where the structural change represents an observable physiological change, here there are discreet and distinct states representing different structures.
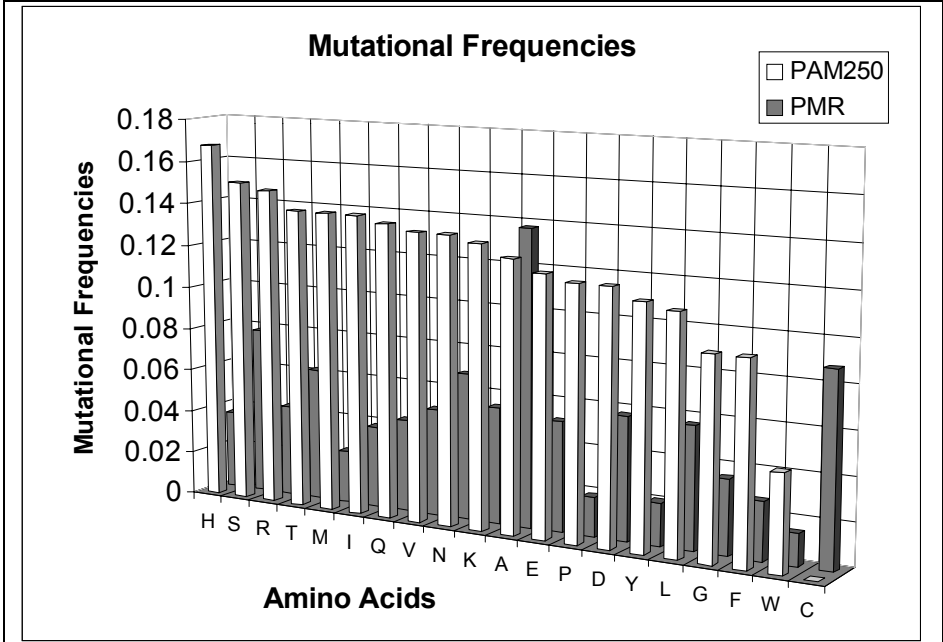
Figure 4. Bar chart of mutational frequencies of amino acids in the PMR and the accepted natural PAM250 amino acid mutation rate [40]. Mutational frequencies for PMR and PAM250 datasets were summed vertically and horizontally (Table 2) for each amino acid to generate a mutational frequency for each amino acid. Mutational frequencies were then sorted by the PAM250 values (foreground).
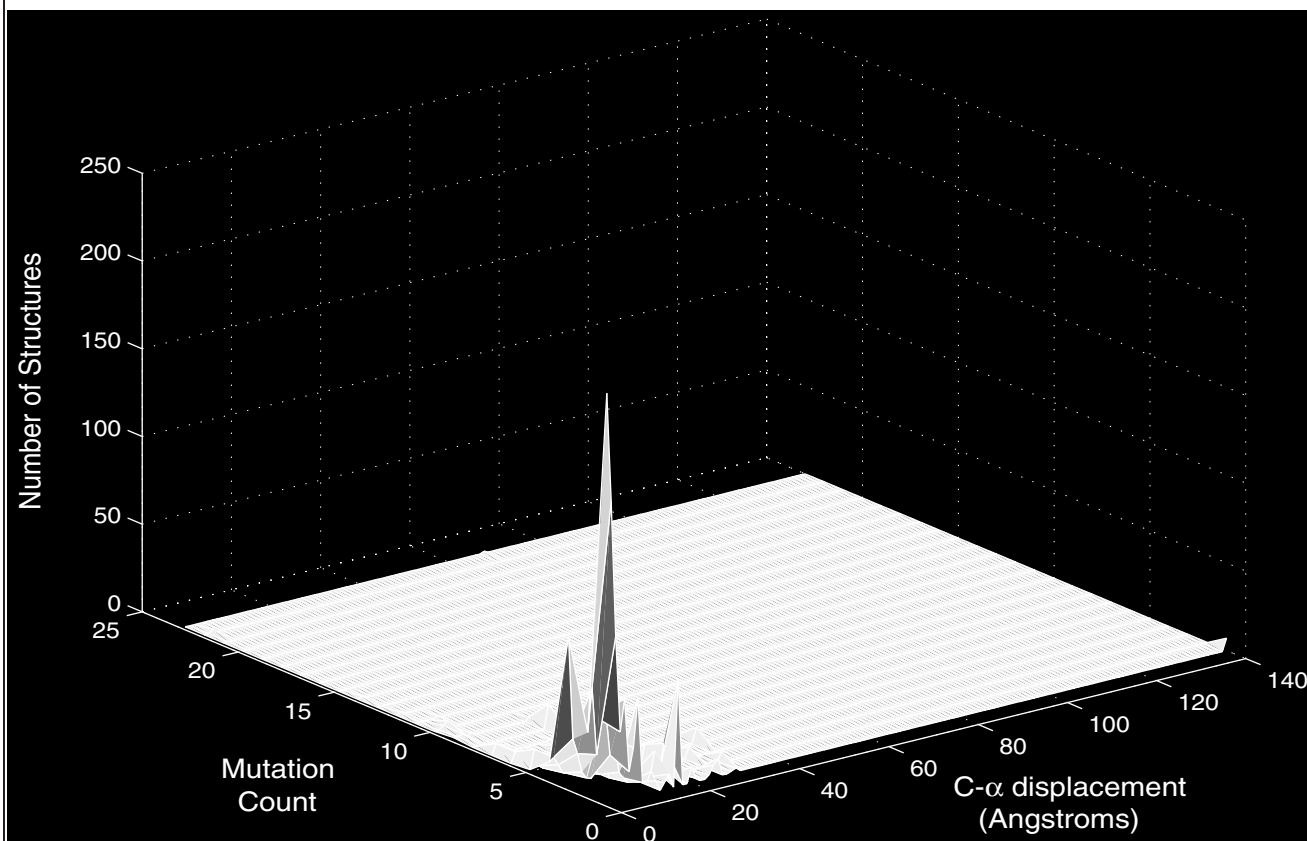
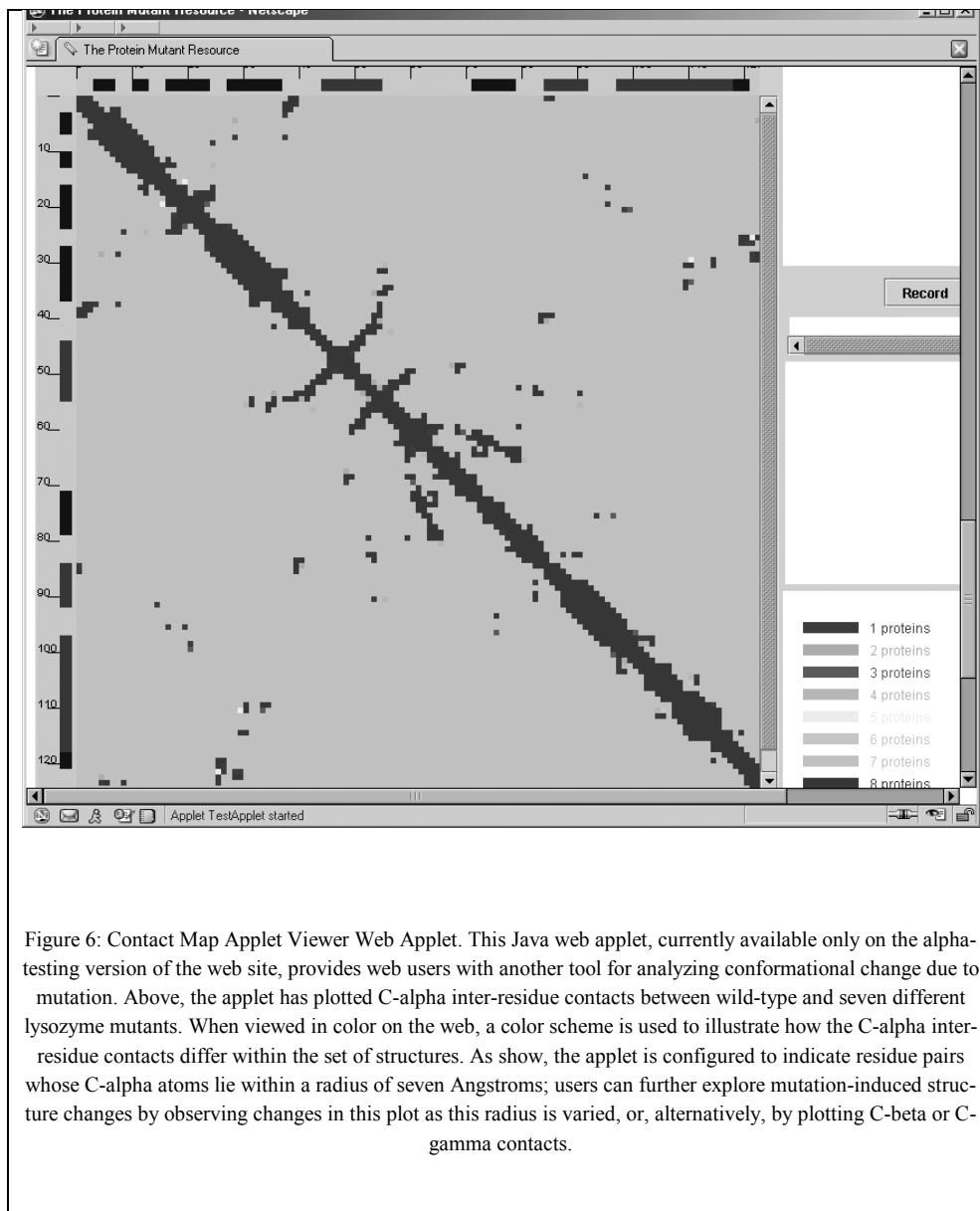**Figure 5. Number of mutations per polypeptide chain versus structural change (C-alpha displacement).**

Figure 6: Contact Map Applet Viewer Web Applet. This Java web applet, currently available only on the alpha-testing version of the web site, provides web users with another tool for analyzing conformational change due to mutation. Above, the applet has plotted C-alpha inter-residue contacts between wild-type and seven different lysozyme mutants. When viewed in color on the web, a color scheme is used to illustrate how the C-alpha inter-residue contacts differ within the set of structures. As show, the applet is configured to indicate residue pairs whose C-alpha atoms lie within a radius of seven Angstroms; users can further explore mutation-induced structure changes by observing changes in this plot as this radius is varied, or, alternatively, by plotting C-beta or C-gamma contacts.
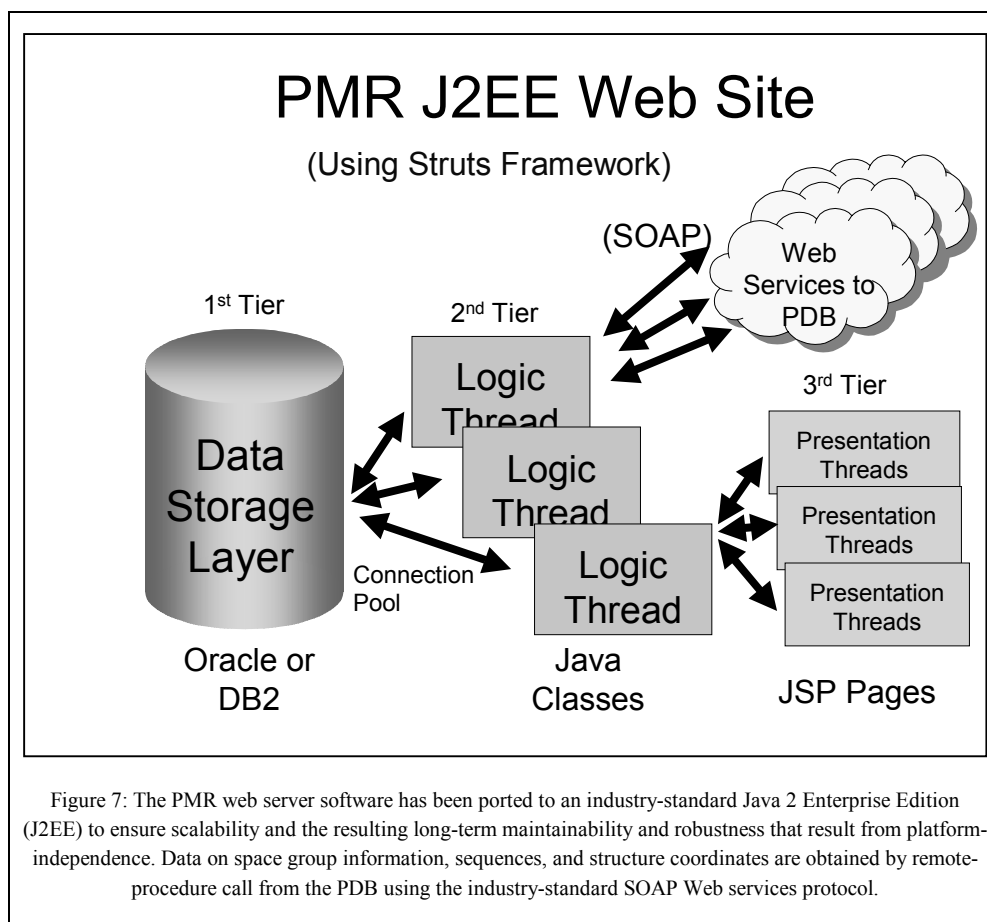
4. Discussion and Conclusion

The addition of a morph analysis to the PMR permits the visualization of conformational change induced by changes to the wildtype protein. However care and a review of the original PDB files is needed to distinguish between conformational change induced by mutation (often indicated by alternative folding, chain breakage, or the passage of back-bone and sidechains atoms through one another) and protein motions caused by ligand

binding rather than mutation, which appear as smooth transitions. It is necessary to often refer back to the original PDB files to understand the cause of conformational change. With improvements to ligand descriptions within the PDB we anticipate better annotating and classifying specific motions in the future.

In general, due to the human eye's remarkable ability to analyze moving objects, morph movies are an exceptional way of comparing two or more similar protein structures. However, the PMR provides several alternatives graphical methods to explore mutation's effect on structure, including a Java-based web applet that dynamically plots, in response to user parameters, maps illustrating the differences in inter-residue carbon contacts between mutant structures (Figure 6).

Induced mutation while retaining structure shows a different patter of substitutions that that observed in nature (Table 2 and Figures 4 and 5). Evolutionary sequence drift usually preserves structure and function whereas structural biologists often mutate proteins specifically to change structure and function, or indeed to induce better structure formation to better nature [44]. A useful addition to the PMR would be mutations that prevented a structure from being observed. Such negative data has not traditionally found its way to the literature or to public databases, but that will likely change with the advent of structural genomics. High-throughput structural genomics studies may also make possible exhaustive mutagenesis structural studies that try to take into account and compensate for the prevailing current biases in structural study mutation frequencies (Table 2), either by adjusting the experimental technique or by statistically messaging experimental results,

The PMR web site provides a web tool to allowing browsing of our extended PDB classifications. We used a new statistical inference method, reported elsewhere [20, 21], to generate a highly accurate new mapping of over 25,000 Protein Data Bank (PDB) sequences to Gene Ontology (GO) terms. Approximately 4,000 new genes were automatically annotated in a highly accurate and statistically rigorous fashion by our algorithm, representing a substantial enhancement of the seed dataset and a classification of more than $2/3^{rds}$ of the PDB. Here, we reported a new, highly efficient, scalable implementation in IBM DB2 SQL and Perl of our original $O(N \lg N)$ algorithm and its iterative application to generate a new dataset covering nearly the entire PDB in a highly-accurate fashion. The statistical scores produced by the method can also be used as a quality control mechanism in purely manual annotation. Our method was able to reduce false positives by more than 10% on PDB data as compared with previous statistical methods in the same class used at similar levels of statistical stringency. We are in the process of applying our method to new databases outside of the PMR and PDB.

PMR J2EE Web Site

(Using Struts Framework)

Figure 7: The PMR web server software has been ported to an industry-standard Java 2 Enterprise Edition (J2EE) to ensure scalability and the resulting long-term maintainability and robustness that result from platform-independence. Data on space group information, sequences, and structure coordinates are obtained by remote-procedure call from the PDB using the industry-standard SOAP Web services protocol.

Our computation of morph movie representations enables the broader structural bioinformatics community to analyze and represent protein mutation conformational change visually. Computation of these biologically interesting results was made tractable by a number of new algorithms [33] and design decisions [27] within our software pipeline. Conversion of the website server software to run within an industry-standard three-tiered platform-independent Java 2 Enterprise Edition (J2EE) framework (Figure 7) ensures greater maintainability and long-term robustness. These results and methods complement existing algorithms [20, 21] used to generate data for other areas of the PMR website.

Acknowledgements

References

[1] A. L. Sabb, G. M. Husbands, J. Tokolics, R. P. Stein, R. P. Tasse, C. A. Boast, J. A. Moyer, and M. Abou-Gharbia, "Discovery of a highly potent, functionally-selective muscarinic M1 agonist, WAY-132983 using rational drug design and receptor modelling," *Bioorg Med Chem Lett*, vol. 9, pp. 1895-900., 1999.

[2] J. M. Hakala and M. Vihinen, "Modelling the structure of the calcitonin gene-related peptide," *Protein Eng*, vol. 7, pp. 1069-75., 1994.

[3] J. H. McKie, "Homology modelling of the dihydrofolate reductase-thymidylate synthase bifunctional enzyme of Leishmania major, a potential target for rational drug design in leishmaniasis," *Drug Des Discov*, vol. 11, pp. 269-88., 1994.

[4] G. Vriend, "WHAT IF: A molecular modeling and drug design program," *J. Mol. Graph.*, vol. 8, pp. 52-56, 1990.

[5] A. R. Fersht, "Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications," *Proc Natl Acad Sci U S A*, vol. 92, pp. 10869-73., 1995.

[6] S. Govindarajan and R. A. Goldstein, "Optimal local propensities for model proteins," *Proteins*, vol. 22, pp. 413-8., 1995.

[7] R. Unger and J. Moult, "Local interactions dominate folding in a simple protein model," *J Mol Biol*, vol. 259, pp. 988-94., 1996.

[8] H. Li, M. Carrion-Vazquez, A. F. Oberhauser, P. E. Marszalek, and J. M. Fernandez, "Point mutations alter the mechanical stability of immunoglobulin modules," *Nat Struct Biol*, vol. 7, pp. 1117-20., 2000.

[9] M. W. Beukers, I. Kristiansen, I. J. AP, and I. Edvardsen, "TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data," *Trends Pharmacol Sci*, vol. 20, pp. 475-7., 1999.

[10] K. Kristiansen, S. G. Dahl, and O. Edvardsen, "A database of mutants and effects of site-directed mutagenesis experiments on G protein-coupled receptors," *Proteins*, vol. 26, pp. 81-94., 1996.

[11] I. Edvardsen and K. Kristiansen, "Computerization of mutant data: the tinyGRAP mutant database," *7TM journal*, vol. 6, pp. 1-6, 1997.

[12] S. M. Maurer, R. B. Firestone, and C. R. Scriver, "Science's neglected legacy," *Nature*, vol. 405, pp. 117-20., 2000.

[13] K. Nishikawa, S. Ishino, H. Takenaka, N. Norioka, T. Hirai, T. Yao, and Y. Seto, "Constructing a protein mutant database," *Protein Eng*, vol. 7, pp. 773, 1994.

[14] T. Kawabata, M. Ota, and K. Nishikawa, "The Protein Mutant Database," *Nucleic Acids Res*, vol. 27, pp. 355-7., 1999.

[15] M. M. Gromiha, H. Uedaira, J. An, S. Selvaraj, P. Prabakaran, and A. Sarai, "ProTherm, Thermodynamic Database for Proteins and Mutants: developments in version 3.0," *Nucleic Acids Res*, vol. 30, pp. 301-2., 2002.

[16] W. G. Krebs, T. M. Nair, and P. E. Bourne, "The Protein Mutant Resource: A Tool for Protein Engineering," *Manuscript in preparation.*, 2003.

[17] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235-242, 2000.

[18] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," *Protein Eng*, vol. 11, pp. 739-47., 1998.

[19] M. Levitt, *STRUCTAL. A structural alignment program*: Stanford University, 1994.

[20] W. G. Krebs and P. E. Bourne, "Statistically Rigorous Automated Gene Annotation and Classification and its Application to Protein Data Bank Sequences using Gene Ontology Terms," *submitted to Bioinformatics*, 2003.

[21] W. G. Krebs and P. E. Bourne, "Statistically Rigorous Automated Gene Annotation and Classification and its Application to Protein Data Bank Sequences using Gene Ontology Terms,". U.S. Patent Pending.

[22] A. G. Rust, E. Mongin, and E. Birney, "Genome annotation techniques: new approaches and challenges," *Drug Discov Today*, vol. 7, pp. S70-6., 2002.

[23] A. L. Bazzan, P. M. Engel, L. F. Schroeder, and S. C. Da Silva, "Automated annotation of keywords for proteins related to mycoplasmataceae using machine learning techniques," *Bioinformatics*, vol. 18 Suppl 2, pp. S35-43., 2002.

[24] M. Biswas, J. F. O'Rourke, E. Camon, G. Fraser, A. Kanapin, Y. Karavidopoulou, P. Kersey, E. Kriventseva, V. Mittard, N. Mulder, I. Phan, F. Servant, and R. Apweiler, "Applications of InterPro in protein annotation and genome analysis," *Brief Bioinform*, vol. 3, pp. 285-95., 2002.

[25] H. Xie, A. Wasserman, Z. Levine, A. Novik, V. Grebinskiy, A. Shoshan, and L. Mintz, "Large-scale protein annotation through gene ontology," *Genome Res*, vol. 12, pp. 785-94., 2002.

[26] W. Krebs, J. Tsai, V. Alexandrov, J. Junker, R. Jansen, and M. Gerstein, "Studying Protein Flexibility in a Statistical Framework: Tools and Databases for Analyzing Structures and Approaches for Mapping this onto Sequences," *Methods Enzymol*, vol. 374, 2003.

[27] W. G. Krebs and M. Gerstein, "The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework," *Nucleic Acids Res*, vol. 28, pp. 1665-1675, 2000.

[28] N. Echols, D. Milburn, and M. Gerstein, "MolMovDB: analysis and visualization of conformational change and structural flexibility," *Nucleic Acids Res*, vol. 31, pp. 478-82., 2003.

[29] M. Gerstein and W. Krebs, "A Database of Macromolecular Movements," *Nucl. Acids Res*, vol. 26, pp. 4280, 1998.

[30] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, 2000.

[31] G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans, "Entrez: molecular biology database and retrieval system," *Methods Enzymol*, vol. 266, pp. 141-62, 1996.

[32] R. A. Laskowski, Hutchinson, E.G., Michie. A.D., Wallace, A.C., Jones, M.L., Thornton, J.M., "PDBsum: a Web-based database of summaries and analyses of all PDB structures," *Trends Biochem. Sci.*, vol. 22, pp. 488-490, 1997.

[33] W. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," *Bioinformatics*, vol. 17, pp. 282-3., 2001.

[34] L. Holm and C. Sander, "Removing near-neighbour redundancy from large protein sequence collections," *Bioinformatics*, vol. 14, pp. 423-9., 1998.

[35] L. Wall, D. Christiansen, and R. Schwartz, *Programming Perl*. Sebastapol, CA: O'Reilly and Associates, 1996.

[36] W. G. Krebs, V. Alexandrov, C. A. Wilson, N. Echols, H. Yu, and M. Gerstein, "Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic," *Proteins*, vol. 48, pp. 682-95., 2002.

[37] M. B. Gerstein, R. Jansen, T. Johnson, B. Park, and W. Krebs, "Studying Macromolecular Motions in a Database Framework: From Structure to Sequence," in *Rigidity theory and applications*, *Fundamental Materials Science*, M. F. Thorpe and P. M. Duxbury, Eds. New York: Kluwer Academic/Plenum press, 1999, pp. 401-442.

[38] E. Martz, "Protein Explorer (software package),".: URL: http://www.umass.edu/microbio/chime/explorer/index.htm, 1999.

[39] A. M. Lesk and C. Chothia, "Mechanisms of Domain Closure in Proteins," *J. Mol. Biol.*, vol. 174, pp. 175-91, 1984.

[40] W. R. Pearson, "Rapid and sensitive sequence comparison with FASTP and FASTA," *Methods Enzymol*, vol. 183, pp. 63-98, 1990.

[41] Y. Kim, S. H. Eom, J. Wang, D. S. Lee, S. W. Suh, and T. A. Steitz, "Crystal structure of Thermus aquaticus DNA polymerase," *Nature*, vol. 376, pp. 612-6., 1995.

[42] B. V. Plapp, H. Eklund, T. A. Jones, and C. I. Branden, "Three-dimensional structure of isoni-cotinimidylated liver alcohol dehydrogenase," *J Biol Chem*, vol. 258, pp. 5537-47., 1983.

[43] H. Li, W. H. Hallows, J. S. Punzi, V. E. Marquez, H. L. Carrell, K. W. Pankiewicz, K. A. Watanabe, and B. M. Goldstein, "Crystallographic studies of two alcohol dehydrogenase-bound analogues of thiazole-4-carboxamide adenine dinucleotide (TAD), the active anabolite of the antitumor agent tiazofurin," *Biochemistry*, vol. 33, pp. 23-32., 1994.

[44] D. W. Heinz and B. W. Matthews, "Rapid crystallization of T4 lysozyme by intermolecular disulfide cross- linking," *Protein Eng*, vol. 7, pp. 301-7., 1994.