Genomics & Informatics Vol. 6(3) 142-146, September 2008

Structural Bioinformatics Analysis of Disease-related Mutations

Seong-Jin Park[¶], Sangho Oh[¶], Daeui Park* and Jong Bhak*

Korean BioInformation Center (KOBIC), KRIBB, Daejeon 305-806, Korea

Abstract

In order to understand the protein functions that are related to disease, it is important to detect the correlation between amino acid mutations and disease. Many mutation studies about disease-related proteins have been carried out through molecular biology techniques, such as vector design, protein engineering, and protein crystallization. However, experimental protein mutation studies are time-consuming, be it *in vivo* or *in vitro*. We therefore performed a bioinformatic analysis of known disease-related mutations and their protein structure changes in order to analyze the correlation between mutation and disease. For this study, we selected 111 diseases that were related to 175 proteins from the PDB database and 710 mutations that were found in the protein structures. The mutations were acquired from the Human Gene Mutation Database (HGMD). We selected point mutations, excluding only insertions or deletions, for detecting structural changes. To detect a structural change by mutation, we analyzed not only the structural properties (distance of pocket and mutation, pocket size, surface size, and stability), but also the physico-chemical properties (weight, instability, isoelectric point (IEP), and GRAVY score) for the 710 mutations. We detected that the distance between the pocket and disease-related mutation lay within 20 Å (98.5%, 700 proteins). We found that there was no significant correlation between structural stability and disease-causing mutations or between hydrophobicity changes and critical mutations. For large-scale mutational analysis of disease-causing mutations, our bioinformatics approach, using 710 structural mutations, called "Structural Mutatomics," can help researchers to detect disease-specific mutations and to understand the biological functions of disease-related proteins.

Keywords: human gene mutation database (HGMD), structural property, physico-chemical property, structural mutatomics

Introduction

Genetic mutations have effects on disease because they alter the function or structure of essential proteins. Protein structure changes by mutation are especially important for understanding the mechanism of diseases that are caused by mutation. Therefore, many mutation studies about disease-related proteins have been conducted using molecular biology techniques in various species. For example, in superoxide dismutase, a known antioxidant, the effect of the L175 mutation causes a decrease in activity, because L175 changes the structural stabilization of the active site (Gabbianelli *et al.*, 1997). Also, in cytochrome P450, known as an important oxidase for drug metabolism, the L358P mutant shows facilitation of electron transfer from the electron donor and acts as a trigger for electron transfer to oxygenated P450. Mutated cytochrome P450 changes from an oxidase to a reductase and causes a loss of function of the oxidase (Tosha *et al.*, 2004). Recently, the I47A/I54V protease mutant in complex with Lopinavir showed that mutation affects the strain of the bound inhibitor in the protease-binding cleft (Grantz Saskova *et al.*, 2008). In previous studies, the mutation of specific sites has been shown to have an effect on the function and structure of proteins that cause disease. It is well known that there is a correlation between mutated proteins and disease. Also, there are bioinformatic tools to predict the correlation between mutation and disease, such as SIFT (Steven Henikoff *et al.*, 2003) and PolyPhen (Vasily Ramensky *et al.*, 2002). However, these tools are based only on sequence homology.

In this study, we conducted a large-scale structural and sequence mutational analysis of amino acids that could have a direct effect on protein function. Because we collected the largest number of 3D structural changes in proteins, such as pockets, we named the dataset the "structural mutatome." The number of such structural mutations will increase continuously, and mapping the mutations to function and to disease will play a critical role in understanding the precise disease mechanisms that are caused by 3D mutations. We classified mutated proteins by their structural properties (distance of pocket residue and mutation, pocket size, surface size, and stability) and physico-chemical properties (weight, instability, isoelectric point, and GRAVY

[¶] Equal contribution

*Corresponding author: E-mail jongbhak@yahoo.com, daeui@kribb.re.kr
Tel +82-42-879-8519, Fax +82-42-879-8535

Accepted 5 September 2008

score). We analyzed the biological meaning of the mutated proteins that were associated with diseases using bioinformatics tools such as Biopython (Chapman *et al.*, 2000), Ligsite (Hendlich *et al.*, 1997), NACCESS (Hubbard *et al.*, 1993), and I-mutant (Capriotti *et al.*, 2004). The overall approach of our study was to map as many structural mutations as possible and find general patterns to analyze 3D mutations with regard to protein function using as many bioinformatic analysis methods as possible. The overall strategy was termed "Structural Mutatomics," because it is intended to find interactions between mutations, structural changes, physico-chemical aberrations, and disease states that are found in the literature and in databases.

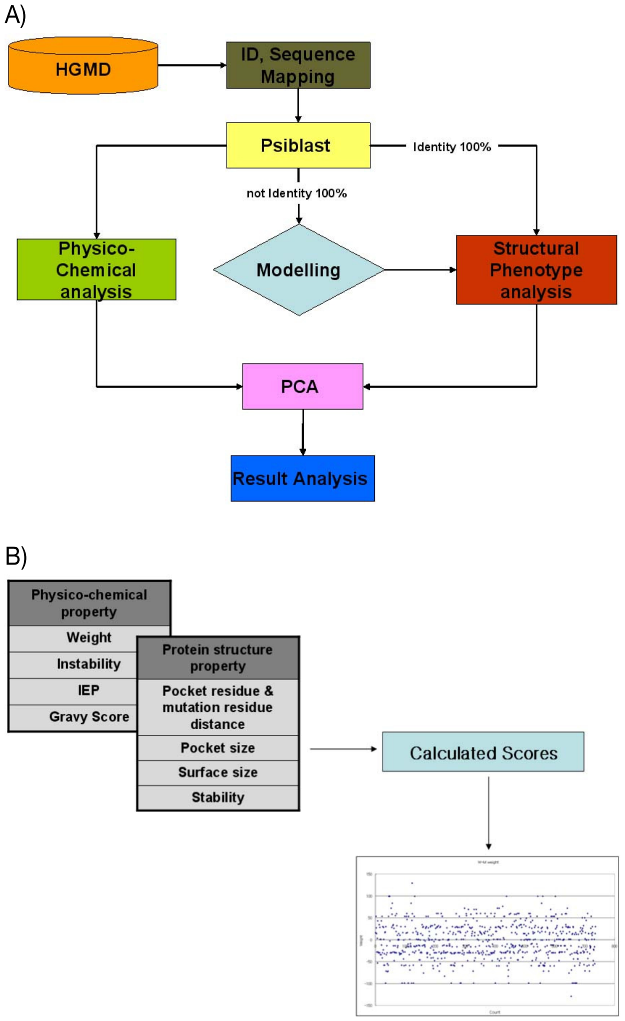

Fig. 1. The flowchart for structural analysis. (A) In order to detect correlations between structure and disease-related mutations, we performed sequential procedures, including conversion of genes to proteins, PSI-Blast, protein structure modeling, physico-chemical analysis, and structural analysis. (B) Scores and Analysis Procedures.

Methods

We constructed a computational pipeline (structural mutatomics pipeline) for structural analysis of mutations that were associated with diseases. Our analysis schema is briefly described in Fig. 1A.

Extraction of mutation information

The Human Gene Mutation Database (HGMD) is a collated database for known (published) gene lesions that are responsible for human inherited diseases. HGMD currently includes information on the nature, location, and sequence context of lesions in human nuclear genes (http://www.hgmd.org; Stensonet *et al.*, 2003). We collected information, including disease, gene symbol, gene name, nucleotide base, codon, OMIM ID, and cDNA accession number from HGMD professional version 2008.1. We retrieved 2899 genes and 43,039 mutations (Table 1), and we stored the information in a local MySQL database (http://www.mysql.com).

Gene-to-protein conversion

In order to map mutations in proteins, we converted the mutated genes into reference proteins. The human protein sequences were retrieved from gene2accession (ftp://ftp.ncbi.nih.gov/gene/) and NCBI (ftp://ftp.ncbi.nih.gov/pub/nrdb/), which have comprehensive, nonredundant, and well-annotated sets. We matched 782 proteins and 5515 mutations from genes in HGMD (Table 1). The match was parsed with a locally developed Perl program (http://www.perl.org).

Protein structure prediction

Proteins with mutations do not always have 3D structures that are solved and deposited in PDB (http://www.rcsb.org/pdb/). Therefore, it is necessary to construct 3D models for many genes. Once we obtain 3D models for any gene, we can look at the location of the mutation

Table 1. The number of genes or proteins and mutations

Name	HGMD	Gene2accession	Modeller
The number of genes or proteins	2,899	782	175
Point mutation	43,039	5,515	710

The final numbers used for the analysis are shown in the last column, under MODELLER program.

in 3D. This is a simple way of detecting what kind of adverse effects that a mutation can have on a protein. For effective structure modeling, 710 human proteins were aligned with proteins that had structures in PDB using the PSI-BLAST (www.ncbi.nlm.nih.gov/BLAST/) algorithm with a cutoff of 30% sequence identity, 70% sequence length coverage, PSI-BLAST iteration 5, and a common expect value (E-value) of 0.0001. If the query protein had 100% sequence identity with any PDB template, the protein was used directly without homology modeling (52 proteins). For 123 proteins, we predicted the 3D structures of the proteins using MODELLER9v4 by the Homology Modeling methodology (John *et al.*, 2003). The MODELLER program automatically constructs an all-atom 3D model using one or more alignments between the query sequence(s) and known homologous structures. We were able to retrieve 175 proteins with mutations out of 710 because not all 782 proteins had homologous structures (Table 1).

Structural analysis

To analyze the correlation between structure and mutation, we calculated the distances between structural pockets on proteins and mutation sites, the change of pocket size due to mutation, the change of protein surface size caused by mutation, and the change of stability affected by mutation. Pockets in proteins are usually critical for their work. Therefore, any nearby mutated pockets can have deleterious effects, causing disease. The distance between a pocket and a mutation was detected as the average RMSD (root mean square distance) of the distances between all residues that participated in a pocket and a mutation residue. To find changes in pocket size, protein surface, and protein stability, we used Ligsite (Hendlich *et al.*, 1997) which calculates the size of pocket means potential ligand-binding site by the PSP (protein-solvent-protein) method, NACCESS (Hubbard *et al.*, 1993) which calculates the atomic accessible area when a probe is rolled around the Van der Waals surface of a macromolecule and I-mutant (Capriotti *et al.*, 2004) which calculates the free energy change of protein stability using a support vector machine. The overall difference between a mutated gene and its wild-type version was measured by Score, calculated as "Wild-type Score - Mutation Score" (Supplemental Table 1, http://www.kogo.or.kr).

Physico-chemical analysis

The physico-chemical properties of mutated proteins are important in order to understand biological functions. We used modules from Biopython (http://biopython.org) (Chapman *et al.*, 2000) to calculate molecular weight, isoelectric point (IEP), protein instability (half life), and GRAVY score (the average hydropathy score for all amino acids) (Park *et al.*, 2008). Also, we calculated the Score-the difference in physico-chemical properties between the wild-type and mutation sequences (Supplemental Table 2, http://www.kogo.or.kr).

Results

To get an intuitive view of each Score, we made a two-dimensional scale plot, assigning the X-axis as mutated protein count and the Y-axis as structural properties or physico-chemical properties (Fig. 1B). The two scores were defined as "Wild-type physico-chemical property scores?Mutated physico-chemical property scores" and "Wild-type protein structural property scores-Mutated protein structural property scores."

The physical distance between protein pockets and mutated residues

To detect which specific mutation patterns were associated with diseases, we retrieved distance scores for the structural phenotypes on two-dimensional scale plots (Fig. 2). Surprisingly, 43% of mutations occurred on residues that were components of pockets. Also, in most cases, the distance was less than 20 Å (confidence 98.5%). This means that the mutated residues were very close to the protein pockets. The mutations that occurred in the pockets of each size in structural models affected the binding and formation of complexes by the mutated proteins. An outlier of this pattern was found for Dilated Cardiomyopathy, a condition in which the

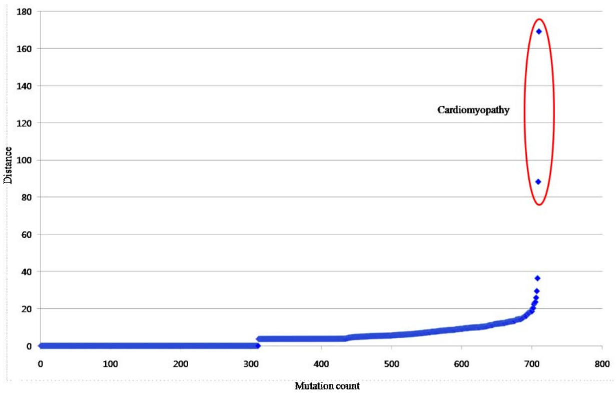

Fig. 2. Two-dimensional analysis of distance. X-axis is mutated protein count. Y-axis is Score of distance. Except for cardiomyopathy proteins, most protein mutations are close to pockets.

heart becomes weakened and enlarged, wherein distance had nothing to do with the mutation. We found that the Dilated Cardiomyopathy protein was very large, and the mutation location was far from the pockets of each size. Most mutations that were found close to the pockets were relatively small in size.

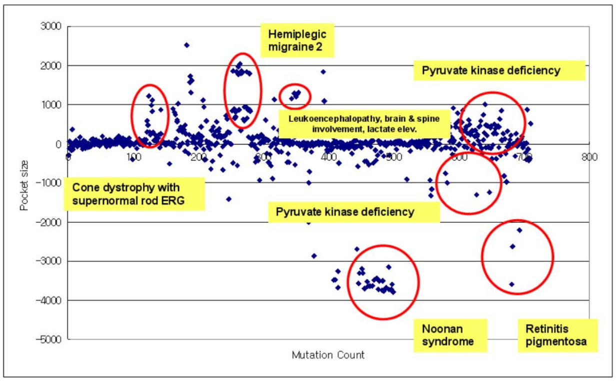

Fig. 3. Two-dimensional analysis of pocket size. X-axis is mutation protein count. Y-axis is score of pocket size. Y-axis shows the pocket size change after mutation, calculated using 3D models.

The difference in pocket size between wild-type and mutated proteins

We calculated the largest pocket size score in a structural phenotype and two-dimensional scale plot (Fig. 3). A mutated protein that is associated with disease usually has an effect on pocket size. The mechanism of protein-protein docking also is affected according to pocket size. In Fig. 3, circled in red, there were pocket sizes that were larger than 2000 Å^{3}. This indicates that a single point mutation that is associated with a disease can be a significant factor in protein structure. For example, the A428T mutation in Impaired Diclofenac Metabolism (NP_000762.2) caused that pocket size to change from 1365 Å^{3} to 1829 Å^{3}, because the mutation created a new hole in the pocket (Fig. 5).

The difference in protein instability index between wild-type and mutated proteins

We also calculated the instability Score for physico-chemical properties in a two-dimensional scale plot (Fig. 4). A mutation site that is associated with disease produces a molecular weight difference in most proteins. Proteins were degraded rapidly above an in-

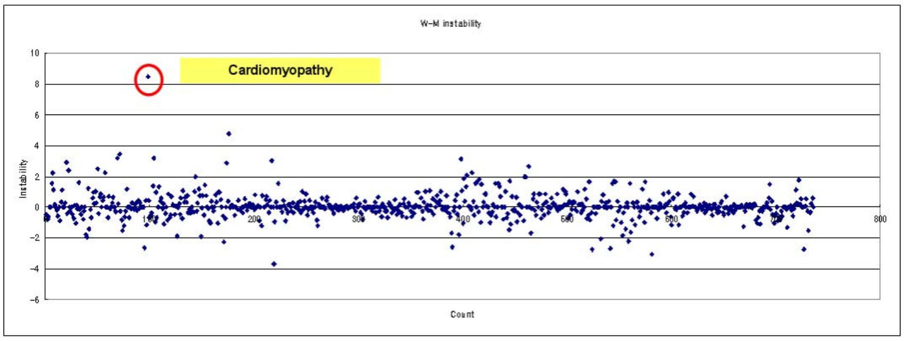

Fig. 4. Two-dimensional analysis of instability. X-axis is mutation protein count. Y-axis is Mutation Score of instability. Most proteins showed relatively small differences before and after mutation in terms of stability.

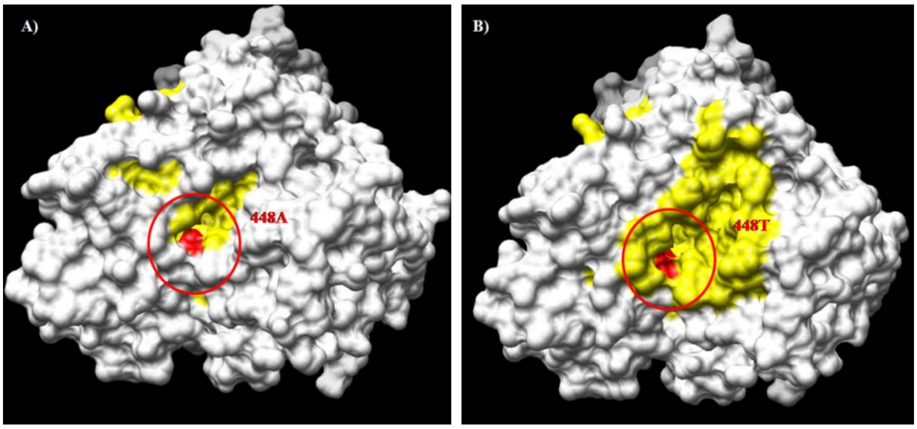

Fig. 5. The change in pocket size by point mutation. A is the 3D structure of a wild-type protein. B is the 3D structure of the A448T mutant. The red surface is a point mutation, and the yellow surface is a pocket site. The pocket size changed from 1365 Å^{3} to 1829 Å^{3} by mutation. The A448T mutation in the NP_000762 protein causes impaired diclofenac metabolism.

stability index of 40. Fig. 4 shows the instability Score. The mutated protein was usually unstable, as seen by an increase in the instability Score. In Fig. 4, the pattern of instability showed values between 2 and -2.

Discussion

For most proteins, the distance between mutated residues and protein surface pockets was less than 20 Å (confidence 98.5%). This means that pockets often are affected by mutations that are associated with diseases. Mutated residues that are close to pockets can change the binding of proteins to other proteins and the formation of complexes. The pocket size also was affected by mutation in our two-dimensional scale analysis of structural property. If a ligand can not interlock with its target protein, it can lead to the failure of protein-protein docking. On the contrary, there was no significant change in stability between the wild-type and mutated proteins in many cases. Our large-scale survey of 3D mutations in the PDB and our models to analyze the effects of mutation on pockets, pocket size, and stability showed that bioinformatic analysis can predict the uncertain effects of mutations on proteins *in vivo* and *in vitro*. This will help researchers to detect more specific mutations and to understand the biological functions of disease-related proteins.

Acknowledgments

We thank our colleagues at KOBIC and Maryana Bhak for editing the manuscript. This work was supported by a grant from the KRIBB Research Initiative Program of Korea, by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MOST) (No. M10508040002-07N0804-00210), and by a Korea Basic Science Institute K-Mep grant (T28021).

References

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33, W306-W310.

Chapman, B., and Chang, J. (2000). Biopython: python tools for computational biology. *ACM SIGBIO Newsletter* 20, 15-19.

Gabbianelli, R., Battistoni, A., Polticelli, F., Meier, B., Schmidt, M., Rotilio, G., and Desideri, A. (1997). Effect of Lys175 mutation on structure function properties of Propionibacterium shermanii superoxide dismutase. *Protein Engineering* 10, 1067-1070.

Grantz Saskova, K., Kozisek, M., Lepsik, M., Brynda, J., Rezacova, P., Vaclavikova, J., Kagan, R., Machala, L., and Konvalinka, J. (2008). Contributing to the Reduced Susceptibility to HIV Protease Enzymatic and Structural Analysis of the I47A Mutation Inhibitor Lopinavir. *Protein Sci.* 17, 1555-1564.

Hendlich, M., Rippmann, F., and Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* 15, 359-363.

Hubbard, S.J., and Thornton, J.M. (1993). NACCESS Computer Program. Dept. of Biochem. and Mol. Biol., University College London.

Jackson, D.A., and Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data. *Environmetrics* 15, 129-139.

John, B., and Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.* 31, 3982-3992.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812-3814.

Park, D.I., Kim, B.C., Cho, S.W., Park, S.J., Choi, J.S., Kim, S.I., Bhak, J., and Lee, S.H. (2008). MassNet: a functional annotation service for protein mass spectrometry data. *Nucleic Acids Research* 36, W491-W495.

Ramensky, V., Bor, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30(17), 3894-3900.

Tosha, T., Yoshioka, S., Ishimori, K., and Morishima, I. (2004). L358P Mutation on Cytochrome P450cam Simulates Structural Changes upon Putidaredoxin Binding. *JBC* 279, 42836-42843.