Gemoetrical verification of protein structure for single nucleotide polymorphism (SNP)

Wonsuhk Uhm[1] **Sung-Geun Lee**[1] **Yang-Seok Kim**[1,2]

[1]*ISTECH Inc., #506 Woongshin Art Plaza, 847 Janghang, Ilsan, Goyang, Gyeonggi, 411-837, Korea*

[2]*Cancer Metastasis Research Center, Yonsei University, College of Medicine, 134 Shinchon, Seodaemun, Seoul, 120-752, Korea*

ABSTRACT: Among non-synonymous SNPs that cause amino acid change in the protein product, the selection of disease-causing SNPs has been of great interest. We present the comparison between the evolutionary (SIFT score) and structural information (binding pocket) to show that the incorporation between them provides an advantage of sorting disease-causing SNPs from normal SNPs. To set up the procedure, we apply the machine learning method to the test data set from the laboratory experiments.

1 INTRODUCION

Human genetic variation is represented by single nucleotide polymorphism (SNP) and many of them are known as the most common type for inherited disease susceptibility in human. However their direct influences to phenotypes are still beneath the full understanding. Especially, non-synonymous SNPs (nsSNP) that cause amino acid change in the protein product are of great interest because it provides the structural and evolutional implications to the protein function and infers the connection to the disease phenotype. Much effort has been devoted to discover such relation based on both protein sequence and structural information.

The algorithms that rely solely on sequence for prediction have been useful tool to determine the phenotypic variation. The hypothesis that the amino acids conserved in the protein homologous family are functionally important and their change leads to deleterious in protein function is presumed. SIFT(Sorting Intolerant From Tolerant)[1] method have demonstrated the use of multiple sequence alignment to identify conserved amino acid site that may be crucial for protein function. SIFT used PSI-BLAST to search against the protein database for homologous sequences and construct a multiple sequence alignment to calculate a position-specific scoring matrix (PSSM). From each matrix entry of probabilities, the phylogenetic entropy is estimated as a balanced average by adding the Dirichlet pseudocounts.

Many other groups have assessed the effect of SNPs in globular protein on the basis of their location in the protein tertiary structure both in PDB and by protein homology modeling [2-5]. Among them, Wang and Moult [2] showed that the most part of disease-causing nsSNP predominantly affects the stability of protein tertiary structure. Also, Sanders and Baker [3] evaluated the structural and evolutionary contributions to deleterious mutation. They fount that a hybrid feature of a solvent-accessibility term and SIFT score obtained the most accurate predictions for deleterious mutation. Stitziel *et al.* [4] showed that the majority of disease-causing nsSNPs from OMIM mapped to potential surface pocket or cavity predicted by AlphaShape algorithm that employs Voronoi diagram and Delaunay tessellation. However, despite of valuable assessment by protein tertiary structure, the lack of protein structure in PDB and inaccuracies of side-chain by protein homology modeling, the accurate verification will remain under future consideration, and therefore it is advisable to use structural information as an assistant to evolutionary analysis. So, by incorporating the suitable structural information with SIFT score, improving the predictive power to determine the deleterious nsSNPs is a main purpose of the present paper.

2 RESULT AND DISCUSSION

Our method was designed to incorporate both the evolutionary and structural information. So, the goal of this work is to build these methods to optimize the predictive power to determine deleterious nsSNP. Another reason to be addressed for this study is to build the automated system to predict the importance of specific residue site for a SNP-related experiment. The experimentalist can focus on the specific important region by this kind of works and, therefore, they can reduce their time and expense.

Total 334 SNPs from 75 OMIM entries were collected to be analyzed for this study. Those protein chains have almost perfect homologous structures at Protein Database (PDB). The position specific scoring matrix (PSSM) for those protein chains was calculated against the database of the non-redundant protein sequence (NR database) provided from National Center for Biotechnology Information. As expected, the entropy values for many disease-causing SNPs from OMIM are distributed at the lower level of the entropy. It is easily understood by considering the fact that the residue site for the diseasing-causing SNPs is evolutionally well conserved, that is to say, low entropy. Figure 1(a) shows clearly this theoretical perspective. But in case of Figure 1(b), there are several SNPs that have relatively higher value of entropy. For example, the residue positions for H101 and R192 score much higher than that of other sites. Note that SIFT scores for these SNPs are classified as "tolerant". One reason for this kind of error is that the number of homologous sequences in protein subfamily to be used for calculation of PSSM is not enough to average. This low homologous sequence issue has been

considered as a weak point of SIFT method. But it will be solved as the database content for those sequences expands. Anther reason can be thought as the fact that those site are not evolutionary conserved but structurally important. So, it is very natural to think the structural information as an assistant of SIFT score.

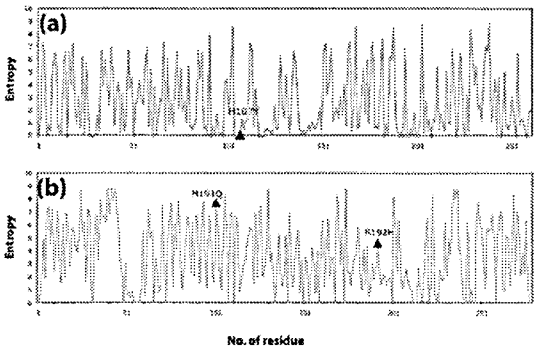

Figure 1 The plot of phylogenetic entropies for (a) OMIM # 259730 (NP_000058) and (b) #600415 (NP_000361). The triangle shows the site for diseasing-causing SNPs.

Actually, in case of Figure 1(b), the homologous subfamily used for calculation is 37 sequences from non-redundant database. So, it is the case for requiring the structural information. We prepare the sequences from PDB to align the protein sequence.

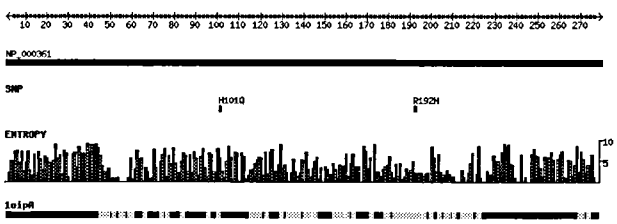

Figure 2 The comparison between the phylogenetic entropy and the location of binding pocket. The light gray regions of lowest bar are corresponds to the residue site which involves the pocket geometry predicted by AlphaShape theory.

Figure 2 shows the comparison between the phylogenetic entropy and the location of binding pocket for OMIM # 600415 used in Figure 1(b). Interestingly, it is clearly shown that two diseasing-causing SNPs are located at the binding pocket.

What we focused on next is to determine the deleterious SNPs from non-synonymous SNPs by incorporating both evolutionary and structural information. Mathematically, it will be a classification problem if there existexperimentally confirmed data set. Here we used the data set from the laboratory experiments (Table 3) compiled by Ng and Henikoff [1] under courtesy of Dr. Ng. So, the general classification methods including Baysian or Support Vector Machine could be applied. Note that there have been many works on this topic, but the exact solution is still under beneath the full understanding. Next step of the data processing is to select the proper features(attributes) for the training the sample data and classifying the test data set. Here, we selected the essential features from generally assumed features (Table 2) by individually measuring information gain with respect to the class by WEKA program [6] and confirmed the selected essential features by the causality rule of Baysian Network.

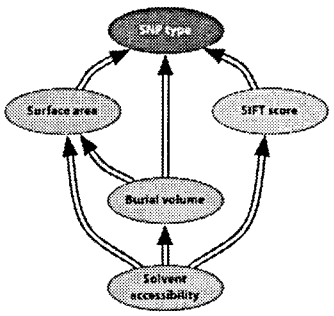

Figure 3 Causality relation among the strongly contributed features.

Table 2 lists the total features used to select the essential features. The mutation type indicate the mutated protein structure after the site-directed mutagenesis on SNP residue, while the wild type means the original protein structure before the mutagenesis. Also, the features listed on a table are basically absolute values of the difference between wild and mutation type.

Figure 3 indicates the causality relation over the essential features. Three features, changes in SIFT score, burial volume, and surface area, influence the SNP type directly, while a change in solvent accessibility indirectly. Interestingly, there is a causality relation between solvent accessibility and SIFT score. Also, we have to note that the suggested essential features are not complete features for classification because these were derived from the laboratory experiments.

With these features, we apply the classification method of Support Vector Machine on a data set of LacI and Lysozyme. Note that although the linear SVM model was used to predict the other methods (Gaussian SVM model) produced the similar result. First, we predict the deleterious SNPs of Lysozyme data set with a training data set from LacI. And then, we apply in opposite data set. Predicting Lysozyme data set ranks very high true positive rate while LacI is confined around 6-70%. This is mainly because the number of deleterious and tolerant SNPs in Lysozyme training data set is not equally balanced.

Data set		Deleterious	Tolerant
Train	Test		
LacI	Lysozyme	155/175 88.6%	1232/1376 89.5%
Lysozyme	LacI	778/1165 66.8%	1648/2267 72.7%

Table 1 Prediction of deleterious and tolerant SNPs in

laboratory experiments data set. LacI and Lysozyme data were used as training and testing data respectively.

The constructed procedure can be applied to classify SNPs from OMIM. But we have to be careful when interpreting the above result. The final goal of this study is to elucidate the method to classify the disease-causing SNPs from non-synonymous SNPs of Human. But, it is still not clear whether the deleterious SNPs will be disease-causing or not. Note that according to the result of Wang and Moult [2] about 80% of SNPs from HGMD database are related with the stability of protein structure.

The lack of protein structure in PDB and inaccuracies of side-chain by protein homology modeling, the accurate verification will remain under future consideration, and therefore it is advisable to use structural information as an assistant to evolutionary analysis.

3 METHOD

We suggest that the location of binding pocket incorporated with SIFT score plays a crucial role in verifying the importance of the SNP site. First, we construct the prediction system to estimate the importance of SNP site. Main contribution for the importance is the phylogenetic entropy and the location of binding pocket. Phylogenetic entropy is the information entropy obtained by summing up the probabilities at each residue site on position specific scoring matrix by SIFT algorithm. So, it can be defined as follows;

$$P_c = -\sum_a p_{ca} \log(p_{ca}), \quad (1)$$

where the summation was done over all amino-acids a at specific site c. And p_{ca} is a probability at each reisidue site and represented as a SIFT score.

The potential surface pocket or cavity are predicted by AlphaShape algorithm that employs Voronoi diagram and Delaunay tessellation. The FOTRAN program "pocket" does all computation step. And we determine a residue site as a pocket site when the total exposed volume to the pocket for all residue atom is greater than 70%.

To calibrate and verify this procedure, we prepare disease-causing nsSNPs from OMIM entry having PDB structures. Total 334 SNPs from 75 OMIM entries were appropriated to this study. We collected these entries into database and made a software to analyze an evolutionary contribution and structural stability by the disease-causing SNPs. Also, additionally, the software provides the annotation information from SWISS-PROT database, secondary structure calculated by Jnet algorithm, and hydropathy by Kyte-Doolittle algorithm. The program was mainly written by PERL with a library of BioPerl. Although those entries are very helpful to describe the importance of specific residue site, we will focus on the phylogenetic entropy and the location of pocket shape in this paper.

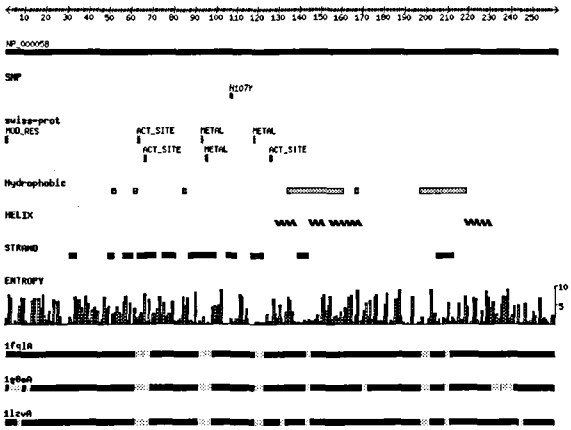

Figure 4 Screenshot of the comparison program for the SNP H107Y of OMIM # 259730(NP_000058).

More nsSNPs are determined as deleterious mutations that have high phylogenetic entropy values or are located near the binding pocket than as reported in Ref. [1, 4].

Interestingly, deleterious SNP sites with the high entropy at the binding pocket have been discovered in many cases. We found that many of those sites correspond to the site with the insufficient evolutionary information such as the case for profiling with few homologous sequences. This result enforces us to believe that incorporating prediction of binding pocket is crucial for insufficient evolutionary information.

Then, we constructed the automated protocol by direct *in silico* mutagenesis to confirm the incorporating power of evolutionary and structural information when determining whether an amino acid substitution in a protein will affect protein function. Structural environment features were mainly derived from the properties of binding pocket and other structural rules described above for the 3D homologous structure.

Type	Properties
Basic	charge, hydrophobicity, polarity
Geometrical	Exposed surface area Exposed pocket area Burial volume Solvent accessibility Total number of pocket Total number of cavity
Chemical	Loss of hydrogen bond Loss of salt bridge Loss of disulfide bond
Evolutionary	SIFT score

Table 2 Selected features to be used to determine the deleterious SNPs from non-synonymous SNPs.

Table 2 lists the features tested in this work. To get the basic and chemical type of features, we applied the rules by

Wang and Moult [2] . The geometrical type of features are drived by applying AlphaShape theory. Solvent accessibility was calculated through DSSP program [7] .

Mutation types from the original protein structures are prepared by the site-directed mutagenesis and generation of side-chain by SCWRL program. Then, we measure the differences of each structural feature between wild and mutation type.

Among those features, crucial ones are carefully selected by evaluating the features individually by measuring information gain with respect to the class and ranking the evaluated features by WEKA program [6] . So, we select top five features shown in Figure 3 which shows the causality relation derived from Bayesian network approach and confirms the relation between the features.

	Deleterious	Tolerant
LacI	1165	2267
Lysozyme	175	1376

Table 3 Data sets from laboratory experiment compiled by Ng and Henikoff [1] .

We applied machine learning methods, support vector machine (SVM), to compare the predictive power as SIFT algorithm or other structure-based algorithms. We used a training and testing set (Table 3 **Data sets from laboratory experiment compiled by Ng and Henikoff [1]** .) based on the laboratory mutation experiments (LacI and lysozyme) [1] compiled by P. Ng.

REFERENCES

[1] P. Ng and S. Henikoff, Predicting deleterious amino acid substitutions, Geome Res 11:864-874, (2001).

[2] Z. Wang and J. Moult, SNPs, protein structure, and disease, Hum Mutat 17:263-70, (2001).

[3] C. Saunders and D. Baker, Evalution of structural and evolutionary contributions to deleterious mutation prediction, J Mol Biol 322:891-901, (2002).

[4] N. Stitziel, T. Binkowski, Y. Tseung, S. Kasif, and J. Liang, topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association, Nucleic Acids Res 32:D520-2, (2004)

[5] S. Sunyaev, V. Ramensky, I. Koch, Lathe, A. Kondrashov, and P. Bork, Prediction of deleterious human alleles, Hum Mol Genet 10:591-7, (2001).

[6] Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[7] W. Kabsch and C. Sander, "Dictionary of protein secondary structure; pattern recognition of hydrogen-bonded and geometrical features", Biopolymers 22:2577-2637, (1983).