BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Prediction model

Volume 6(1)

An ANN model for the identification of deleterious nsSNPs in tumor suppressor genes

Vinod Chandra[1]*, Rejimoan Ramakrishnan[2], Shalini Ramanathan[2]

[1]Department of Computer Applications, College of Engineering Trivandrum, Kerala, India; [2]Department of Computer Science, P.S.G. College of Technology, Coimbatore, Tamil Nadu, India; Vinod Chandra - Email: vinodchandrass@gmail.com; Phone: 91 471 2515531; Fax: 91 471 2598370; *Corresponding author

Received February 07, 2011; Accepted February 17, 2011; Published March 02, 2011

Abstract:
Human genetic variations primarily result from single nucleotide polymorphisms (SNPs) that occurs approximately every 1000 bases in the overall human population. The non-synonymous SNPs (nsSNPs), lead to amino acid changes in the protein product may account for nearly half of the known genetic variations linked to inherited human diseases and cancer. One of the main problems of medical genetics today is to identify nsSNPs that underlie disease-related phenotypes in humans. An attempt was made to develop a new approach to predict such nsSNPs. This would enhance our understanding of genetic diseases and helps to predict the disease. We detect nsSNPs and all possible and reliable alleles by ANN, a soft computing model using potential SNP information. Reliable nsSNPs are identified, based on the reconstructed alleles and on sequence redundancy. The model gives good results with mean specificity (95.85%), sensitivity (97.40%) and accuracy (96.25%). Our results indicate that ANNs can serve as a useful method to analyze quantitative effect of nsSNPs on protein function and would be useful for large-scale analysis of genomic nsSNP data.

Keywords: SNP, nsSNP, ANN, Tumor suppressor genes

Availability: http://www.snp.mirworks.in

Background:
Single Nucleotide Polymorphism (SNP) represents the most abundant class of genetic variations in the human genome. Non-synonymous SNPs (nsSNPs), which cause the changes of amino acid residues in proteins, account for almost half of all DNA mutations and may be functionally neutral or deleterious **[1, 2]**. The disease causing variations may cause deleterious effects on proteins. They may inactivate the functional sites or interaction sites of enzymes or impact the folding of proteins and may significantly destabilize the stability of proteins, or change the solubility of proteins **[3, 4]**. So these variations represent critical molecular markers for dissecting the biological mechanisms underlying complex diseases, as well as for Pharmacogenomic studies. Such markers have become very popular for all kinds of genetic analysis, and disease like cancer. Cancer suppressors and Oncogenes play an important role in the control of the cell cycle, apoptosis, angiogenesis, and development processes that are under pressure of purifying selection **[5]**. Therefore, protein-damaging mutations in cancer-related genes would be expected to be under the pressure of purifying selection and thus to have a lower population frequency. The relationships between the genotype and phenotype of nsSNPs in tumor suppressor genes have received a plenty of research attentions because of their prevalence in the drug responses and cancer therapy **[6, 7]**. Tumor suppressor genes are normal genes that slow down cell division, repair DNA mistakes, and apoptosis or programmed cell death. Non synonymous SNPs (nsSNPs) are the main cause of these mutations and to mining nsSNPs from cancer related genes considered as a laborious process and done only by site directed mutagenesis experiments and gene knock out/knock in experiments. Recently, some groups have tried to evaluate the deleterious nsSNPs based on 3-dimensional (3D) structure information of proteins and homology based SIFT (Sorting Intolerant from Tolerant) algorithm (http://sift.jcvi.org/, **[1]**). Some other methods based on site entropy calculations, relative stability changes were also developed for predicting deleterious nsSNPs **[8]**. These methods based on protein sequence have been demonstrated that the accuracy is the same as other methods using tertiary structure information. However, the theoretical prediction methods for deleterious nsSNPs are still in its infancy since the 3D structural information of most proteins are still unavailable. To overcome these, our primary challenge is that how to accurately predict those potentially deleterious nsSNPs. Deleterious nsSNPs prediction for the tumor suppressor genes has received great focus from experimental researchers. In this work, we suggest a computational model used to predict deleterious nsSNPs in tumor suppressor genes. Evolutionary conservation features and changes in the physicochemical properties of amino acid are used as parameters for ANN. The method predicts deleterious SNPs from a dbSNP id or a SNP sequence. Both fasta and raw formats are acceptable as input sequence. ANN verification is done for predicted deleterious SNPs. A database search is also included for known deleterious nsSNPs.

Methodology:
Datasets:
Details regarding the genes were collected from NCBI entrez genes (http://www.ncbi.nlm.nih.gov/gene). The protein sequences were obtained from Swiss-Prot database (http://expasy.org/sprot/) and NCBI human genome protein sequence (http://www.ncbi.nlm.nih.gov/). The databases of Swiss-Prot

sequence variants provide full information of classification about nsSNPs associated with the given gene. The variants are labeled as disease, polymorphism or unclassified. The training dataset contains (1) deleterious variations dataset: 68 nsSNPs were collected from seven tumor suppressor genes **(Table 1 see Supplementary material)**. Deleterious variations were labeled as disease in the Swiss-Prot database and (2) neutral variations dataset: 124 nsSNPs were collected from seven tumor suppressor genes **(Table 1)**. Neutral variations were labeled as disease in the Swiss-Prot database. A test dataset was also prepared by the same method for analyzing the performance measures of our prediction system.

Evolutionary conservation feature:
Evolutionary conservation score check whether a substitution of any other amino acid is tolerant or intolerant in every position especially the variant site **[9]**. The evolutionary conservation features are calculated by Position specific score in matrix value (PSSM) for amino acid substitution, multiple sequence alignment of the protein sequence with its homologous sequence, compute the frequency of variant at the mutation site and pseudo-counts and Dirichlet distribution densities are used due to inadequate sequence diversity. Position specific value of residue type calculation is given in **Supplementary material**.

Multiple sequence alignment of the homologous sequence is calculated by ClustalW (http://www.ebi.ac.uk/Tools/msa/ clustalw2/). Due to the limitation of homologous sequences, a normal score calculation is not efficient. So, pseudo count and Dirichlet mixture methods are adapted to increase the results. For pseudo-count calculation, the numbers of pseudo sequences are added with similar properties of the available sequences. Dirichlet distribution is often used as the prior distribution in Bayesian analysis of data where each observation is one of the limited set of possibilities **[9]**. Our case, 20 amino acids constitute a set and each Dirichlet density describes a particular residue composition that might occur at an alignment column. The Dirichlet mixture is a linear combination of these densities with multiplying coefficient of their occurrence **[10]**. Therefore, sum of the coefficient is given as 1. A set of densities and mixture coefficients has been fitted to the BLOCKS database of protein multiple alignments. In this work, we have used component Dirichlet mixture derived from BLOCKS database (http://blocks. fhcrc.org/, **[11]**).

Physicochemical properties:
In this study many properties of amino acids are considered and filtered out based on the relevance of our prediction. Finally, following physicochemical properties of amino acid sequence are considered. They are, changes in hydrophobicity, free energy of transfer from inside to outside of a globular protein, hybridization potential based on free energy of transfer (kcal/mol), hydrophilicity, optimized matching hydrophobicity, molecular weight of amino acids, average flexibility index and solvent accessibility of amino acids. From the parameters input matrix are created for the prediction system.

ANN prediction:
The prediction of the deleterious SNP is carried out with an adaptive artificial neural network (ANN). In this study, ANN was chosen as the tool for the prediction, as they are powerful classifiers whose ability to cope with complex data and their potential for modeling data of high non-linearity **[12, 13]**. We used a feed forward Multi Layer Perceptron (MLP), with four layers. From the selected twelve parameters (four evolutionary conservation and eight physic-chemical properties), undoubtedly the input layer of the neural network must have twelve neurons. In the hidden layers, various combinations were tried out and we got the best results as eight neurons in the first hidden layer and four in the second hidden layer, by trial and error. The output layer has one neuron. Back propagation algorithm was used to train the network. Training can be performed with use of several optimization schemes and there is access to exact partial derivatives of network outputs versus its inputs. The learning rate and momentum were initially set at 0.2 and 0.8 respectively. The training dataset is divided into two subsets. First subset (70% of the total training data) was used to train the neural network. Second subset was used to stop the training process once the model had reached the performance conditions like optimal error value thus preventing over training. Once the training is stopped, the efficiency of the model was further assessed by presenting another data subset, to determine the performance for unseen cases which were not involved in the training process **[14]**. Optimization was done by repeating the process with different data subsets. The optimization needs nearly 500 epochs for this network.

Discussion:
A vast numbers of SNPs are seen in human chromosome and presently there are more than a million SNPs in dbSNP that can be screened for association with diseases. We used a good dataset for human nsSNP predictions from the Swiss-Prot annotated 'disease' and 'polymorphism' variants of known human proteins. Variants annotated neutral polymorphisms may have an unknown association with disease. For prediction type problems, a prediction can be either positive or negative. These counts falls into four categories: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). These contents are used to calculate sensitivity (true positive rates), specificity (1-false positive rates) and total prediction accuracy for assessment of the prediction system. A test dataset is prepared which consists of protein sequences, obtained from the Swiss-Prot database and NCBI human genome protein sequence. The mean specificity (95.85%), sensitivity (97.40%) and accuracy (96.25%) of the ANN prediction were obtained from the test results. Using novel combination of parameters, the ANN predictor performs well in terms of specificity, sensitivity and accuracy. When compared with the previous theoretical studies, Ridge Partial Least Square (RPLS) and Linear Discriminant Analysis (LDA), our method has better performance measures. The prediction accuracy of the RPLS model and LDA model are 84.8% and 80.4% respectively **[1]**. This improvement in accuracy of our model is due to combined feature framework into an ANN predictor. Our method uses amino acid sequence properties and other theoretical prediction methods for deleterious nsSNP prediction requires 3D structural information of proteins, but 3D structural information of most of the proteins are still unavailable.

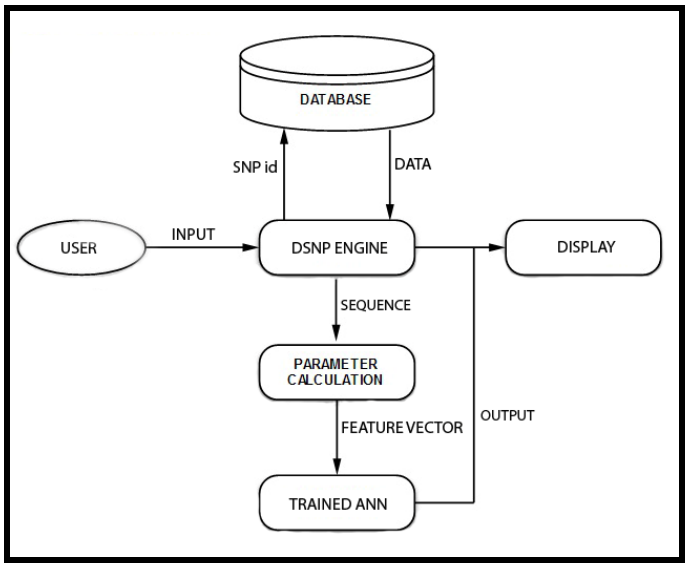

Figure 1: Organizational chart of deleterious nsSNP prediction system. The input to the method is in two forms: dbSNP id and amino acid sequence. If the sequence is dbSNP id, search the database for similar entries. The output contains general information of the gene and nsSNP, information on the variant and sequence based information.

We investigated the role of two parameter sets (evolutionary features and physicochemical properties) in the ANN predictor. A measure of how each feature set contributes to the prediction performance of ANN can be calculated in the course of training. We estimated the relative importance of two feature sets in prediction. For that purpose, input layer of the ANN was redesigned by removing one feature set from the input vector field. The performance of the ANN was evaluated by giving a new feature vector after omitting a feature set whose performance being evaluated. Same dataset is used for training and evaluation of the ANN. The test result is compared with mean accuracy of the ANN predictor. The accuracy decreased to 23.4% and 17.5% from the original ANN predictor's accuracy. Physicochemical attributes gives the information sufficient to identify the amino acid substitution involved. But results are better when evolutionary features and physicochemical properties are combined together. This demonstrates that each parameter in the feature set makes a significant positive contribution to the overall performance of the model, though predictability. Thus, any good predictor of result which relies upon a single set characteristic will fall short of the accuracy obtainable by a combination of characteristics.

The system architecture of the prediction system is given in **Figure 1**. In this system, a database search and a prediction model are incorporated. Two forms of input (dbSNP id and amino acid sequence) are acceptable for the prediction system. For dbSNP id inputs, database search is carried out. For an amino acid sequence input (in fasta / raw format), after removing the invalid characters, calculate the parameters. These values are given as the input of the ANN predictor. The prediction model is deployed into a web server, which finds

utility of nsSNP identification and it may leads to tumor studies for scientific community. The database search and a prediction models are incorporated in the web server. The interface of the deleterious nsSNP web server is shown in **Figure 2**.

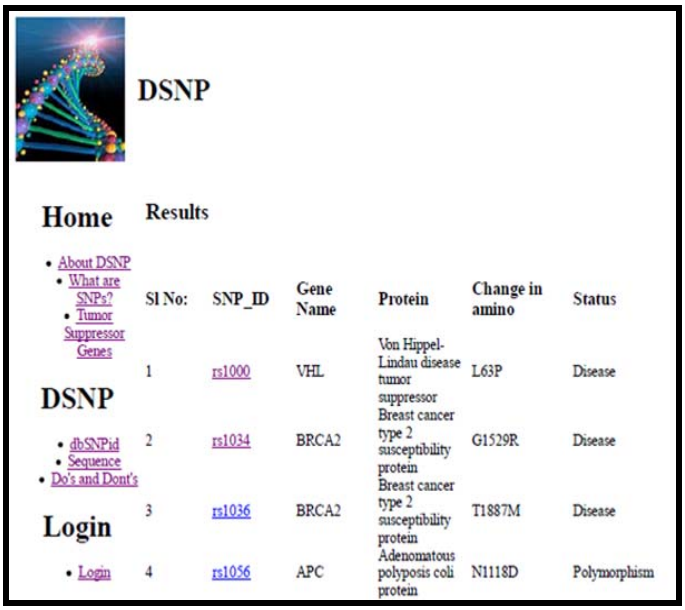

Figure 2: The interface for prediction and results. SNP_IDs are available at NCBI site.

Conclusion:

The functional analysis of deleterious nsSNPs may offer lead to understand the genetic differences between individuals and disease states. It eventually improves the medical treatments by allowing the prediction of genetically related disease risk and drug response. Therefore, computational prediction of deleterious nsSNPs gives great importance and promotes the development of pharmacogenetics and pharmacogenomics. At protein structure level, in-vitro experimental data have shown that the deleterious nsSNPs might contribute to changes of the structure, stability and function of proteins. At amino acid sequence level, the deleterious nsSNPs might be expected to produce a least conservative replacement. In this work, parameters of position-specific features and scale of amino acid change were calculated and is used to predict deleterious nsSNPs in the tumor suppressor genes using ANN, combining physicochemical properties of amino acids. Available genes and nsSNPs are included in the database search.

References:

[1] Li Y *et al. FEBS Lett.* 2006 **580**: 6800 [PMID: 17141228]
[2] Care MA *et al. Bioinformatics* 2007 **23**: 664 [PMID: 17234639]
[3] Pei J & Grishin NV. *Bioinformatics* 2001 **17**: 700 [PMID: 11524371]
[4] Ng PC & Henikoff S. *Genome Res.* 2006 **12**: 436 [PMID: 11875032]
[5] Wang Z & Moult J. *Hum Mutat.* 2001 **17**: 263 [PMID: 11295823]
[6] Bao L & Cui Y. *FEBS Lett.* 2006 **580**: 1231 [PMID: 16442527]
[7] Sunyaev S *et al. Trends Genet.* 2000 **16**: 198 [PMID: 10782110]
[8] Dobson RJ *et al. BMC Bioinformatics* 2006 **7**: 217 [PMID: 16630345]
[9] Zimmerman JM *et al. J Theor Biol.* 1968 **21**: 170 [PMID: 5700434]
[10] Babak & Radford. *Journal of Machine Learning Research* 2009 **10**: 1829
[11] Pietrokovski S *et al. Nucleic Acids Res.* 1996 **24**: 197 [PMID: 8594578]
[12] Khan J *et al. Nat Med.* 2001 **7**: 673 [PMID: 11385503]
[13] Chan HL *et al. Bioinformatics* 2005 **21**: 2191 [PMID: 15746277]
[14] Chandra V *et al. BMC Bioinformatics* 2010 **11** Suppl 1: S2 [PMID: 20122191]

Edited by P Kangueane
Citation: Chandra *et al.* Bioinformation 6(1): 41-44 (2011)

BIOINFORMATION

open access

Supplementary material:

Position specific value of residue type 'a' at position 'u' is given by

$$q_{u,a} = \frac{\alpha f_{u,a} + \beta g_{u,a}}{\alpha + \beta}$$

$$g_{u,a} = \sum f_{u,b} * p_a$$

$$f_{u,a} = n_{u,b} / N_{seq}$$

Where,

α , β - scaling parameters.
$f_{u,a}$ - frequency of residue type a at position u.
$g_{u,a}$ - number of pseudo-counts of residue type 'a'.
p_a - probability of residue 'a' occurring at any position in any sequence.
$n_{u,b}$ - number of residue 'b' at position 'u'
N_{seq} - number of sequences in multiple sequence alignment

Table 1: Dataset used for ANN training. 68 deleterious SNPs and 124 non-synonymous deleterious SNPs are attached to the corresponding Tumor Suppressor Genes collected from Swiss-Port and NCBI human genome protein sequences.

Tumor Suppressor Genes	Deleterious Members	Neutral Members
APC	11	9
BRCA1	14	27
BRCA2	17	34
MEN1	14	2
PTEN	10	1
TP53	2	25
VHL	0	26