

Neuro-Fuzzy Approach to Classification of Human Non-Synonymous SNPs Based upon Statistical Geometry

Maxim Barenboim, Iosif I. Vaisman, and D. Curtis Jamison

Keywords: non-synonymous SNP, neural networks, fuzzy logic, common disease, Delaunay tessellation

1 Introduction.

The ability to predict the effect of non-synonymous SNPs (nsSNPs) on protein function is important for the success of genetic disease association studies. Simple Mendelian disease mutations are observed with very low frequency in the population and can be traced through family pedigrees. It is much more challenging to identify nsSNPs contributing to polygenic diseases. According to the common disease-common variants hypothesis, common diseases are caused by both genetic and environmental factors and affect much larger percentage of population. Direct association analysis examines those nsSNPs which have a high probability of not only being associated with disease, but which also may directly contribute to disease. Computational assessment of the impact of nsSNPs upon protein function is an important tool for association studies. Current computational tools use artificial intelligence methods such as decision trees, support vector machine or neural networks to assess the impact of nsSNPs, and provide researchers with binary Boolean logic terms as the results. This is not always appropriate for classification of functional effects of nsSNPs, where the functional effects of polymorphisms on the protein might be somewhere between two extremes: either the protein is completely functional and the nsSNP is neutral, or the mutation completely disrupts a protein function. In order to account for the spectrum of possible impact, we have applied fuzzy logic to the assessment of nsSNPs. Using datasets of disease and neutral amino acid variants having known 3D structure and using characteristics as attributes obtained by statistical geometry technique based on Delaunay tessellation, we prioritize unclassified nsSNPs by applying a neuro-fuzzy schema. Classification of unclassified nsSNP set in terms of disease potential is available at <http://rna.gmu.edu/FuzzySnps/>

2 Methods.

Neuro-fuzzy schema. To prototype the neuro-fuzzy system, the Matlab Neural Network (NN) and Fuzzy Logic (FL) Toolboxes were used (Figure 1A) [1]. The first component is a back-propagation NN with 17 input neurons, 5 neurons in the hidden layer, and 2 output neurons. The FL component consists of five layers (Figure 1A): an input layer, fuzzification layer with six membership functions, layer with nine fuzzy rules, output layer with five membership functions (Figure 1B), and a centroid defuzzification layer.

Training dataset and dataset of unclassified nsSNPs. For our nsSNP data set, we used three different datasets of human variants extracted from Swiss-Prot variant web pages [2]. The disease-associated dataset (daSNPs) consisted of substitutions annotated as “disease,” which refers to disease-causing mutations with less than 1% frequency in a population and to disease-associated polymorphisms found in at least 1% of a population and associated with disease in medical literature and databases. Based on classification methodology most of the nsSNPs associated with complex diseases are assumed to be in the daSNP set. The neutral nsSNP dataset (ntSNPs) consisted of nsSNPs tagged as “polymorphism”, which are presumably neutral amino acid substitutions not affecting protein functions and without known links to disease. These two datasets are used for training and validation of neuro-fuzzy system. The third, unclassified dataset (unSNPs) refers to variants that have not been categorized yet and most likely consists of a mixture of disease-associated and neutral nsSNPs. The error of the neuro-fuzzy system was estimated with stratified tenfold cross-validation.

Attributes. As attributes for NN input neurons, we used characteristics obtained by applying a novel statistical geometry technique based on Delaunay tessellation to separate daSNPs from ntSNPs. In this technique, the 3D space of the protein is mapped as tetrahedrons using the alpha-carbons as the vertices. The main characteristic of Delaunay tessellation is that the number of nearest neighbors in 3D space is at all times four [3]. We identify an objective set of characteristics which differentiate daSNPs from ntSNPs, such as difference in total potential (ΔQ), volume, and tetrahedrality. Clustering amino acid substitutions to conservative and non-conservative groups and using a 3-letter alphabet based on side chain polarity shows significantly lower ΔQ in non-conservative changes to daSNPs than when hydrophobic residues were substituted by charged or by polar residues. The daSNPs in the protein core also cause much lower ΔQ than surface daSNPs [4]. Each categorical value was converted to a separate input, continuous data were normalized to range between 0 and 1.

3 Results.

The NN gives two crisp outputs for disease risks associated with nsSNP. The FL component takes these two inputs and returns a single disease potential of nsSNP on the scale from 0 to 100, and provides with the linguistic determination of output membership (Figure 1B). The accuracy of the FL approach was much better than that of the straight NN, as estimated through the stratified tenfold cross-validation (Table 1). Classification of the unSNP set shows that of the 568 nsSNPs, 413 were classified as SNPs with high disease potential in the range from 92 to 50; 117 nsSNPs classified as medium with disease potential ranging from 50 to 34; and 38 classified as low disease potential SNPs scoring from 28 to 21.

4 Discussion.

The advantage of NN is the ability to be trained with unprocessed data. However, NN is a “black box” in terms of the rules leading to its output. On the other hand, FL systems are transparent since they are built upon explicit IF-THEN rules, but are unable to learn and adapt to changing conditions [1], [5]. The merger of NN with FL yields a system that can learn and can be amenable to human perception [5], [6]. In case of nsSNPs, we show that the FL approach built upon rules derived from statistical geometry leads to a marked improvement in the accuracy of prediction for disease alleles. The error rate associated with prediction of neutral SNPs is still high, as these SNPs might be as yet unidentified contributors to polygenic diseases. Moreover, we are currently unable to define precise rules for each possible situation, since a particular SNP could belong to both the disease and neutral classes to some degree, depending on the presence of other SNPs in the genome and external environmental factors. Thus, the FL approach seems quite appropriate, compared e.g. to a Bayesian approach where the uncertainty is related to the likelihood of the event outcome. The FL approach allows us to assess the disease potential of nsSNPs and to select the most promising nsSNPs for further investigation. Inclusion of attributes such as degree of amino acid conservation and structural rules in order to increase the sensitivity of this inference system is currently underway.

Table 1: Accuracy of NN and FL components of inference system.

Component	daSNP error %	ntSNP error %	Total error %
NN	16	44	25
FL	2	44	15

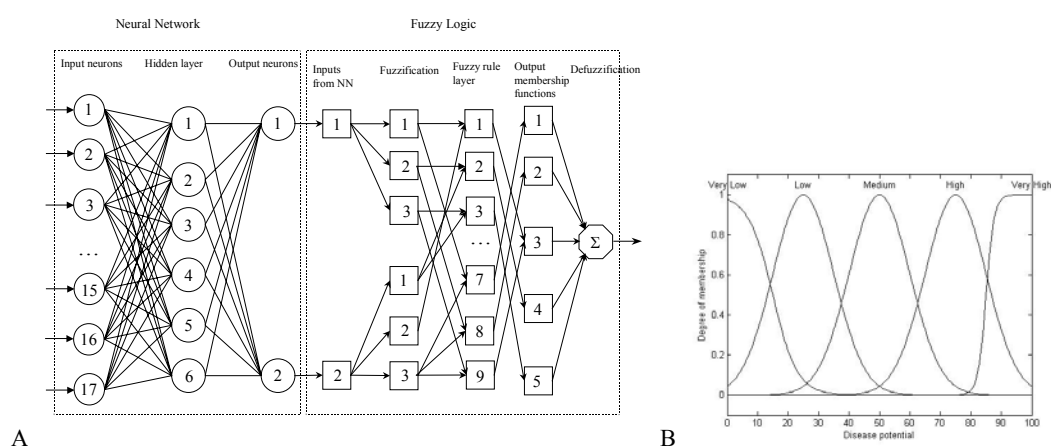


Figure 1: (A) Neuro-fuzzy schema for evaluating the disease potential of nsSNPs; (B) Output membership functions.

References

[4] Barenboim, M., Jamison, D.C. and Vaisman, I.I. 2005. A Computational Geometry Approach to the Study of Functional Effects of Human Non-synonymous SNPs. *submitted*.

[5] Goonatilake, S., and Sukhdev, S. 1995. *Intelligent Hybrid Systems*. Toronto: Wiley Publishing.

[6] Nauck, D., Klawonn, F. and Kruse, R. 1997. *Foundations of Neuro-Fuzzy Systems*. New York: John Wiley.

[1] Negnevitsky, M. 2005. *Artificial Intelligence: A Guide to Intelligent Systems*. Addison-Wesley.

[3] Singh, R.K., Tropsha, A. and Vaisman, I.I. 1996. Delaunay tessellation of proteins: four body nearest neighbor propensities of amino acid residues. *Journal of Computational Biology* 3: 213-21.

[2] Yip, Y.L., Scheib, H., Diemand, A.V., Gattiker, A., Famiglietti, L.M., Gasteiger, E. and Bairoch, A. 2004. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Human Mutation* 23:464-70.