PREDICTING DELETERIOUS SINGLE NUCLEOTIDE POLYMORPHISMS

Petr Jaša
Doctoral Degree Programme (1), FIT BUT
E-mail: ijasa@fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@fit.vutbr.cz

ABSTRACT

Human genetic variations are primarily the result of single nucleotide polymorphisms (SNPs) that occur approximately every 1000 bases in the overall human population. In this paper we focus on non-synonymous protein coding single nucleotide polymorphisms (nsSNPs) which can affect protein structure or function. Because of that, nsSNPs are believed to have the largest impact on human health compared with SNPs in other regions (non-protein coding) of the genome. It is also very important to distinguish those nsSNPs that affect protein function from those that are functionally neutral. This article provides an introduction to SNP problematic and an overview of recent used data sources and bioinformatics methods for prediction deleterious nsSNPs which play significant role in human susceptibility to diseases or drugs.

1. INTRODUCTION

The recent sequencing of the human genome has revealed a wealth of information concretely several million genetic variations between individuals. There has been great expectation that the knowledge of an individual's genotype will provide a basis for assessing susceptibility to diseases and designing individualized therapy. It has been estimated that 90% of genetic variations in humans are due to single nucleotide polymorphisms (SNP) [6]. Within the genome we can distinguish, as it is shown in table 2, several types of SNPs. Non-synonymous single nucleotide polymorphisms (nsSNPs) that lead to an amino acid change in the protein product are of particular interest because they account for nearly half of the known genetic variations related to human inherited diseases.

Some nsSNPs are linked to a disease condition but others are not related with any change in phenotype and so they are regarded as neutral. Several studies [2, 3, 4, 6, 7] have attempted to predict the functional consequences of a nsSNP, namely whether it is disease related or neutral, based on attributes of the polymorphism. Some attributes depend only on the sequence information, for example the types of residue found at the SNP location. Structural attributes such as solvent accessibility can be chosen if the protein sequence containing the nsSNP has a known 3D structure or is highly similar to a protein sequence of known structure. To facilitate the identification of disease-associated nsSNPs from a large number of neutral nsSNPs, it is important to develop computational tools to predict

the phenotypic effects of nsSNPs. Such tools are recently based on empirical rules or machine learning.

The first part of this article briefly describes biological background of single nucleotide polymorphisms and predicting of their dangerousness. The next part introduces basic principles used in predicting effects of nsSNPs to human health and shows predictors which are being used within the predictions. Next, this paper provides the survey of existing amino acid substitutions (AAS) prediction methods and discusses their performance and usability. Finally is presented our intention in this field for the future term.

2. BIOLOGICAL BACKGROUND

In this section are described biological notions that are mentioned in this article. There is no need to know these notions in all their scope, thus this part is rather in the form of glossary of terms.

DNA	It is nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. The main role of DNA molecules is the long-term storage of information.
Genome	It is the whole hereditary information encoded in the DNA. It includes both the genes and the non-coding sequences of the DNA
Gene	It is section of DNA with specific function. Gene can encode protein or RNA or can have regulation function.
Protein	It is large organic compounds made of amino acids arranged in a linear chain. The sequence of amino acids in a protein is defined by a gene. Order of amino acids determines the 3D structure of protein.

Table 1: Glossary of terms [10]

Coding SNPs	cSNPs	Positions that fall within the coding regions of genes
Regulatory SNPs	rSNPs	Positions that fall in regulatory regions of genes
Synonymous SNPs	sSNPs	Positions in exons that do not change the codon to substitute an amino acid
Non-synonymous SNPs	nsSNPs	Positions that incur an amino acid substitution
Intronic SNPs	iSNPs	Positions that fall within introns

Table 2: SNP functional classes [1]

3. PRINCIPALS OF PREDICTING EFFECTS SNP TO PROTEIN

Recent studies have discovered a variety of potential predictors discriminating disease-associated nsSNPs from neutral nsSNPs. Empirical rule-based and machine learning approaches were used to classify these two types of nsSNPs. Prediction rules discriminating

disease-associated and neutral nsSNPs were derived based on structural information, evolutionary information [4] or both [9]

Methods based on evolutionary information use multiple sequence alignment of relative sequences. Then it is possible to find out the position where some changes occur during the evolution. The observation that disease-causing mutations are more likely to occur at positions that are conserved throughout evolution, as compared with positions that are not conserved, suggested that prediction could be based on sequence homology [3]. The prediction accuracy depends heavily on the existence of a sufficient number of homologous sequences. Saunders and Baker (2002) showed that the prediction accuracy decreased significantly when fewer than 5–10 homologous sequences are available. In such cases is crucial to incorporate some other information. It was also observed that disease-causing AASs had common structural features that distinguished them from neutral substitutions. Thus, structure information could also be used for prediction [3].

3.1. SEQUENCED-BASED PREDICTION METHODS

Sequence-based AAS prediction methods take an input sequence and search it against a sequence database to find homologous sequences. A multiple sequence alignment of the homologous sequences reveals what positions have been conserved throughout evolutionary time, and these positions are inferred to be important for function. The AAS prediction method then scores the AAS based on the amino acids appearing in the multiple alignment and the severity of the amino acid change. An amino acid that is not present at the substitution site in the multiple alignment can still be predicted to be tolerated if there are amino acids with similar physiochemical properties present in the alignment. For example, if a protein sequence alignment shows aromatic acid like tyrosines and tryptophans at a particular position, one would expect that the other aromatic amino acid, phenylalanine, would also be tolerated at that position. [3]

3.2. STRUCTURAL-BASED PREDICTION METHODS

Structure-based AAS prediction methods accept an input sequence and find the best match against a protein structure database. Because most structure-based AAS prediction methods use general structural features surrounding the site of substitution and do not require detailed information at the atomic level, they can model the substitution onto the structure of a homologous protein. They do not require the exact structure of the input sequence. Then AAS prediction methods determine the position of the AAS and based on that can take into account several structure factors such as carbon-beta density, solvent accessibility, crystallographic B-factor, and the difference in free energy between the new and the old amino acid. [3]

3.3. ANNOTATIONS-BASED PREDICTION METHODS

AAS prediction methods can also include some annotations into analysis to make prediction more accurate. The Swiss-Prot database annotates the positions of a protein that are located in the active site, are involved in ligand binding, are part of a disulfide bridge, or are involved in other protein-protein interactions [3]. For example, if the position of the AAS is annotated as involved in ligand binding, then the AAS is predicted to affect the protein.

4. BIOINFORMATICS TOOLS FOR PREDICTING DELETERIOUS SNPS

Because experimentally analysis of the impact of each nsSNP on the protein structure and potentially on its function would be extremely time consuming and expensive it is desirable to use some computational methods for this task. There are few problems in starting to widely use these methods. The main problem is that recent data sources contain incomplete information. For some sequences is known their structure for some sequences is not. Even if we know the structure there is no evidence that the function annotations are known. Because of that the coverage of particularly method is decreased. Next problem is that data is spread in many particular databases. SNP information is currently collected in several databases [1], including: dbSNP, the Human Genome Variation Database (HGVbase), SWISSPROT, the Japanese Single Nucleotide Polymorphism (JSNP) database and the HapMap Project and more. If we want to know sequence and structure data and also, for example, pathway relations between proteins we must integrate the information from some of these databases. Finally there is insufficient amount of information about nsSNPs which are responsible for deleterious changes in human phenotype. Then the training sets are imbalanced and the accuracy of predicting declines.

On the other hand, currently used methods are quite robust and are able to come up to above mentioned problems. Currently the state-of-the-art classification tools are based on support vector machines SVMs or decision trees and the best features for classification are just based on structural and evolutionary properties. Today are available several web resources, where the tools with the best results are for example Polyphen [9], SIFT [4], topoSNP, SNAP or stSNP. PolyPhen uses the features based on sequence, evolutionary and structurally information. But not all features are obligatory. SIFT (sorting intolerant from tolerant) is available online for predicting intolerant mutations using position-specific information derived from sequence alignments, and requires only sequence and homologue information.

	PolyPhen	SIFT	topoSNP	SNPs3D
Input	protein sequence and AAS, dbSNP id, HGVBase id, or protein id	protein sequence, AAS, dbSNP id or protein id	protein id, or protein sequence	dbSNP id, protein id, literature search, or gene onthology
Output	score from 0 to positive number, where 0 is neutral and a high positive number is damaging	score from 0 – 1. 0 means damaging, 1 neutral	location of substitution on protein and conservation reported separately	score <0 is damaging, mutation on protein structure can be vizualized

Table 3: SNP functional classes

In the table 3 is provided a brief description of the most widely known resources for SNP analysis. It is only reference table, there isn't enough place for detailed description of all parameters in this paper. In addition, all resources are currently evolving and maybe they have already better parameters now.

5. CONCLUSION AND FUTURE PLANS

The prediction of deleterious effects of non-synonymous single nucleotide polymorphism on human health is believed to be very useful and strong tool for assignment appropriate individual treatments or preclude inherited diseases. Each individual person can have different reaction to given drug according to distribution of SNPs in its genome. If we were able to determine deleterious SNP biasing the effect of given drug we would simple choose such drug that is best for the particular person.

Many new studies and projects in the field of analysis of SNPs and always increasing amount of genomic data are not provided easy survey for us. Thus our attention will be any further concentrated on detailed exploration of recent methods and resources. According to the result we will focused either to developing own general method for predicting deleterious nsSNPs or to improving existing methods for special group of protein. Both possibilities includes preparing suitable set of SNPs for training and testing methods based on learning machine approach, finding of appropriate predictors and also finding profitable combination of known methods.

REFERENCES

[1] Mooney, S.: Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis, 2004

[2] Donson, J. R. et al.: Predicting deleterious nsSNPs: an analysis of sequence and structural attributes, 2006

[3] NG, C. P Henikoff, S.: Predicting the Effects of Amino Acid Substitutions on Protein Function, 2006

[4] NG, C. P Henikoff, S.: SIFT: predicting amino acid changes that affect protein function, 2003

[5] Burke, F. D. et al.: Genome bioinformatic analysis of nonsynonymous SNPs, 2007

[6] Bao, L. Cui, Y.: Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information, 2005

[7] Tian, J. et al.: Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines, 2007

[8] Bromberg, Y. Rost, B.: SNAP: predict effect of non-synonymous polymorphisms on function, 2007

[9] Polyphen, [http://genetics.bwh.harvard.edu/pph]

[10] Wikipedie, [http://en.wikipedia.org]