Mathematical Statistics
Stockholm University
Examensarbete **2012:4**,
http://www.math.su.se/matstat

Statistical Modeling of the Severity of Mutations from Protein and Genetic Data

Anna Olofsson[*]

May 2012

Abstract

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation in humans. The number of SNPs identified in the human genome is growing rapidly, but attaining experimental knowledge about possible disease-associated variants is a laborious quest, and the main challenge is to narrow the list down to a few candidate genes where the mutations occur. At the moment the identification of candidate genes is quite intuitive. Current in-silico, mathematical and statistical tools provide only a very basic, sequence-based indication about the relevance of a mutation to a disease. There is a lack of multifactorial tools applying statistical, mathematical and biological knowledge to automatically estimate how interesting or relevant a mutation is to a disease by scoring it in some appropriate way: The higher the score the more likely it is that the mutation is disease-causing.

In this paper three SCoring Methods (SCM1-SCM3) are created for estimating the relevance of a mutation to a disease, separating deleterious mutations from neutral ones, each based on two types of data sets. The first one is PolyPhen-2, a web-based software tool, estimating the probability of a possible impact of a mutation on the protein level. The second one is the 1000 Genomes Project, an online catalogue storing information about variations in the population (i.e., the allele frequency). These two factors are combined in different ways for the investigated scores.

Either p-values were calculated, using training data and Fisher's exact and combined tests (SMC1), or logistic regression was used for predicting the probability that a mutation is harmful (SMC2), or a linear combination of the two factors was used as score (SMC3). In order to quantify how well benign mutations are separated from harmful ones, we used the area under the receiver operating characteristic curve, AUC.

[*]Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: annaaolofsson@gmail.com. Supervisor: Ola Hössjer.