BMC Bioinformatics

BioMed Central

Oral presentation

Open Access

Ranking single nucleotide polymorphisms by potential deleterious effects

Phil Hyoun Lee* and Hagit Shatkay

Address: School of Computing, Queen's University, Kingston, Ontario, Canada, K7L 3N6

Email: Phil Hyoun Lee* - lee@cs.queensu.ca

* Corresponding author

from Fourth International Society for Computational Biology (ISCB) Student Council Symposium
Toronto, Canada. 18 July 2008

Published: 30 October 2008

BMC Bioinformatics 2008, **9**(Suppl 10):O6 doi:10.1186/1471-2105-9-S10-O6

This abstract is available from: http://www.biomedcentral.com/1471-2105/9/S10/O6

Introduction

Identifying single nucleotide polymorphisms (SNPs) that are responsible for common and complex diseases such as cancer is of major interest in current molecular epidemiology. However, due to the tremendous number of SNPs on the human genome, to expedite genotyping and analysis, there is a clear need to prioritize SNPs according to their potentially deleterious effects to human health. As of yet, there have been few efforts to quantitatively assess the possible deleterious effects of SNPs for effective genetic variation studies. Here we propose a new integrative scoring system for prioritizing SNPs based on their possible deleterious effects in a probabilistic framework.

Methods

We aim to quantitatively measure the potential *deleterious* effects of SNPs on four bio-molecular function of their genomic region, namely, *splicing*, *transcription*, *translation*, and *post-translation modification*. Figure 1 outlines the three main steps of our assessment process.

For simplicity, we refer to the assessed score as the *functional significance* (FS) *score* of a SNP.

STEP 1. Retrieving Predicted Functional Information

Given a set of SNPs, we first retrieve their predicted functional categories (i.e., 'deleterious' or 'neutral') and corresponding confidence scores for the decisions (i.e., $S \in R$) using 16 publicly available web-services and databases. The confidence scores are then normalized onto the common scale.

STEP 2. Computing Tool Reliability

We define the Tool Reliability (TR) score as how likely each tool is to correctly predict deleterious SNPs, and estimate it based on the tendency of the tool to make consistent predictions with others, following the approach proposed by Long and his colleagues [1].

STEP 3. Computing Functional Significance

Given the prediction results and normalized confidence scores obtained in step I and the TR score computed in step II, the FS score of a SNP is computed as the average of the normalized confidence scores, weighted by the reliability of each tool, as summarized in Figure 1.

Conclusion

We applied our method to 126,496 SNPs located in 607 disease-susceptible genes obtained from the OMIM [2] database (downloaded Jan. 2008). The assessment results show that splice sites and coding regions are most enriched with SNPs with highly putative deleterious effects, which is consistent with previous findings about functional SNPs [3]. We further validated our scoring system by checking out that the distribution of the FS scores for SNPs known to be disease-causing is significantly different from that of SNPs selected uniformly at random within the same gene (p-value 1.0303e-055, paired t-test, $\alpha = 0.05$).

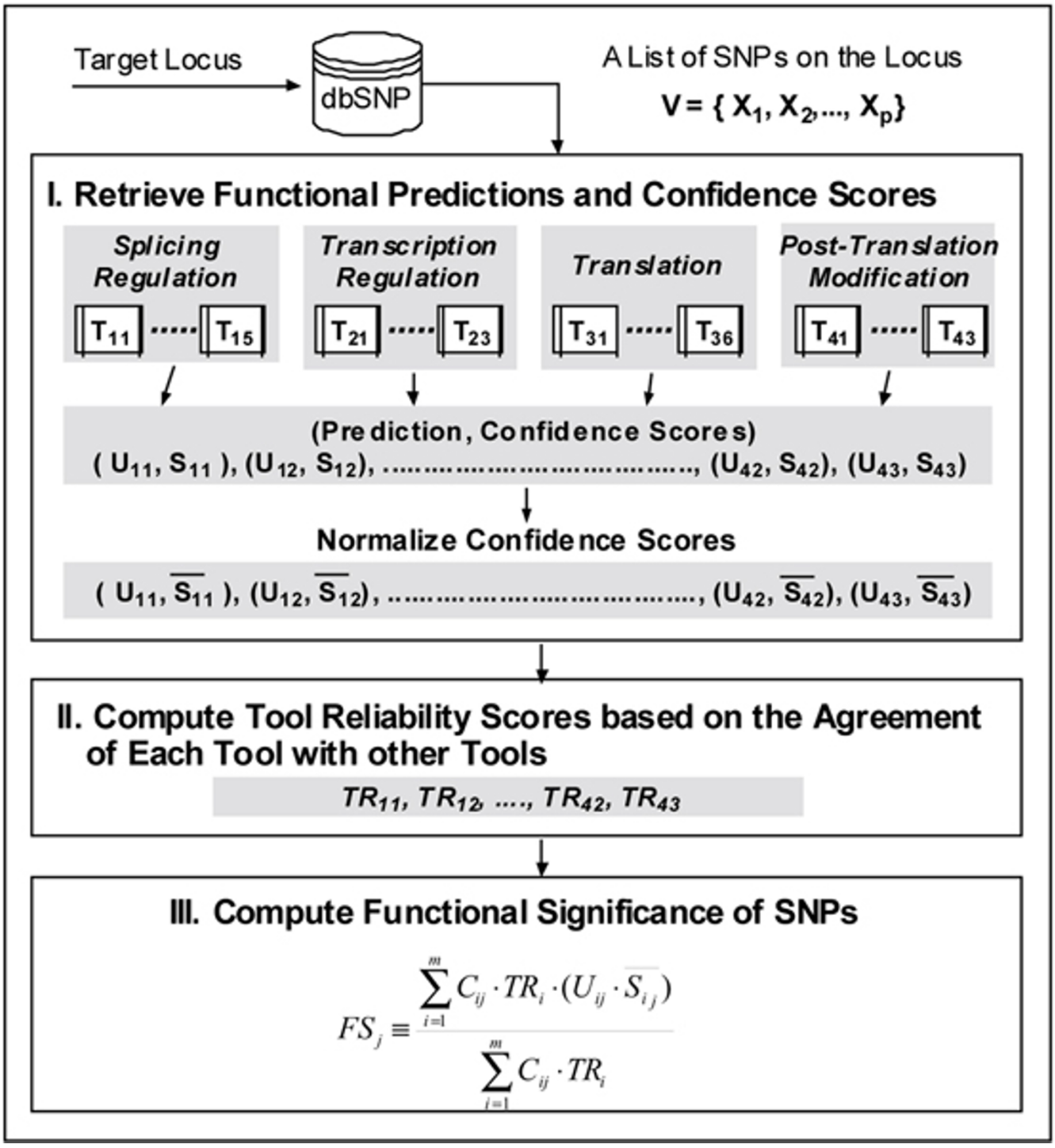

Figure 1
Outline of our assessment process.

References

1. Long PM, Varadan V, Gilman S, Treshock M, Servedio RA: **Unsupervised evidence integration.** *Proceedings of the 22nd international Conference on Machine Learning: 7–11 August 2005; Bonn, Germany* 2005:521-528.
2. **Online Mendelian Inheritance in Man, OMIM™** [http://www.ncbi.nlm.nih.gov/omim/]
3. Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, Vance JM, Zuchner S, Hauser MA: **SNPselector: a web tool for selecting SNPs for genetic association studies.** *Bioinformatics* 2005, **21(22):**4181-4186.