Prediction of regulatory sequence variations

Virginie Bernard, Patrick Tan, David J. Arenillas, Wyeth W. Wasserman

The convergence of high-throughput technologies for sequencing individual full-genomes, and rapid advances in genome annotation are driving a neo-revolution in human genetics. This wave of family-based genetics analysis is revealing causal variations. By mapping millions of short DNA "reads" to the human genome reference a list of small nucleotide variations, insertions and deletions is obtained. Selecting variations shared by relatives having the same disorder reveals causal candidates. Further analyses are required to reveal those variations more likely to contribute to a disease phenotype. Existing software scores the severity of changes based on amino acid changes, but do not consider variations outside of protein encoding regions. Such variation may be deleterious for the regulation of gene expression. While protein-encoding exons occupy 2% of the genome, the remaining 98% of the genome controls the developmental and physiological profile of gene activity - when and where a gene will be active. Non-coding regions are therefore of high interest. Functional contributions of cis-regulatory sequence variations to genetic disease are numerous. The need for bioinformatics methods to identify regulatory variations is imperative.

Given full-genome sequence data, we can predict regulatory variations. Our software system enables genetics researchers to prioritize variations damaging for the gene regulation. As a first step, by analyzing all variations, inside and outside exons, we predict those more likely to alter splice sites or transcription factor binding sites (TFBSs). In order to predict variations impacting TFBSs, we use position weight matrices available from reference databases or derived from experimental archives of protein-DNA interactions. Focusing on variations within regulatory regions derived from ChIP-seq data improves reliability. In order to predict variations impacting splicing recognition, we used splice site predictor software. For both cases, TFBS and splice site, we focused on variations leading to the loss or the gain of a predicted site. Bioinformatics methods for identification of regulatory site alterations will be increasingly important with advance in genome analysis.