Inferring SNP Function Using Evolutionary, Structural and Computational Methods: Session Introduction

M. Dimmic, S. Sunyaev, and C. Bustamante

Pacific Symposium on Biocomputing 10:382-384(2005)

INFERRING SNP FUNCTION USING EVOLUTIONARY, STRUCTURAL, AND COMPUTATIONAL METHODS

MATTHEW W. DIMMIC
Dept. of Biological Statistics and Computational Biology, Cornell University
Ithaca, NY 14853, USA

SHAMIL SUNYAEV
Dept. of Medicine, Brigham & Women's Hospital, Harvard Medical School
Cambridge, MA 02115, USA

CARLOS D. BUSTAMANTE
Dept. of Biological Statistics and Computational Biology, Cornell University
Ithaca, NY 14853, USA

Single nucleotide polymorphisms (SNPs) are the most prevalent form of genetic variation within populations. Recent technological advances have enabled the accumulation of massive amounts of data on SNPs—more than 15 million entries in dbSNP alone—from within a range of species (e.g., human, *Drosophila*, *Anopheles*, mouse, dog, *Arabidopsis*, maize, and *Plasmodium*). Efficient and accurate prediction of a mutation's effect promises to accelerate research in a myriad of fields, ranging from medicine and agriculture to basic genetics and evolutionary biology.

Functional SNPs affect the structure or function of DNA, RNA, or proteins. The effect on the molecular function is in some cases translated further into an effect on the organism phenotype. If the phenotypic effect impacts survival and reproduction, natural selection operates on the SNP alleles. Consequently, evidence of the functional importance of SNP variants can come from three different sources. Structural biology and biochemistry can detect the influence of amino acid or nucleotide substitutions at the molecular level; association (linkage) studies pursue the detection of correlation among SNP variants with specific phenotypes of interest; evolutionary and population genetics detects natural selection by means of statistical analysis. Bioinformatics makes possible a correlated study of the bulk of recent data on human SNPs through a variety of computational approaches, which include ideas from all of the above fields. The papers in this session represent a diversity of statistical and

computational advances towards elucidating the evolutionary and biophysical functional significance of particular SNPs.

The frequency of a SNP is related to both the rate of mutation and the selective forces acting on that mutation; the relative contributions of each process can affect analysis of a SNP's functional impact. To tease apart mutation bias from natural selection, Yampolsky and Stolztfus employ a novel measure of amino acid exchangeability (EX), which is based on the results of thousands of mutagenesis experiments on protein activity. Applying this measure to both between-species and within-species variation, they find that the apparent contribution can change quite drastically with increasing evolutionary distance. Their analysis on hominid variation, for example, finds a sharp cutoff point in the relationship between fixation probability and amino acid exchangeability.

Presumably, amino acid exchangeabilities are governed by the interplay of physiochemical features of amino acids, location in protein structure, and molecular function. Methods which explicitly account for these features should aid in predicting which SNPs will have an effect on protein function. In their contribution to this session, Karchin and co-workers use mutual information to measure which changes in amino acid features correlate most strongly with *in vivo* functional effects. When used as inputs in a support vector machine (SVM) classifier, the most mutually informative measures were also better at predicting which SNPs were most likely to affect function in several different protein families.

The degree of selection acting on functional SNPs is difficult to calculate without an accurate estimate of the underlying mutation rate, which can vary widely across the genome. In their session paper, Rogozin and co-workers focus on the biochemical mechanisms of the nucleotide mutational process at mutational hotspots. Examining a variety of mechanisms, they find that a large proportion of mutations can be explained by oxidative damage on the nucleotide level. An analysis of mutations in human mitochondrial genes finds that the molecular mechanism of mutation differs significantly between hypervariable and coding regions, hypothesizing that this difference may be caused by dislocation mutagenesis.

The paper by Webb-Robertson *et al.* describes a Bayesian formulation to compare the power of molecular evolutionary models to predict the distribution of observed SNPs. Their analysis indicates that simple models which treat only mutational effects perform roughly as well as the current models of amino acid exchangeability, evidence that there is still a great deal of room for improvement on current methodologies.

These identification and modeling methods assume that an organism's genetic variation has already been described and that a set of SNPs is already

available. One promising technique for assembling the map of an individual's genetic variation is optical mapping, where fragments of DNA are bound to a surface, cleaved, and visualized using light microscopy. Anantharaman and co-workers describe an algorithm for reassembling the ordering of these fragments based on the location of the cleavage sites. They demonstrate how this algorithm can be used to infer the parental haplotypes of a diploid organism, an advance which holds great promise for the genome-wide study of how variation is maintained in species over generational time.

Acknowledgments

The session organizers would like to thank all those who submitted manuscripts to this session, and we are grateful to the anonymous referees for their careful reviews.