

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

DOE/ER/62970-1

KERNEL PRINCIPLE COMPONENT ANALYSIS OF MICROARRAY DATA

Final Report
for Period September 1, 2000 – August 31, 2002

Fatemeh Haghighi

Columbia University
New York, New York 10032

November 2003

Prepared for

THE U.S. DEPARTMENT OF ENERGY
AWARD NO. DE-FG02-00ER62970

DOE Patent Clearance Granted
MP Dvorscak
Mark P Dvorscak
(630) 252-2393
E-mail mark.dvorscak@ch.doe.gov
Office of Intellectual Property Law
DOE Chicago Operations Office
11.20.03
Date

The following describes the research conducted during the course of the DOE funding period. Given the limitations in the quantity and quality of gene expression array data, the focus of the research was shifted to development of statistical and computational tools for evaluation and detection of disease susceptibility mutations within a large set of individuals or even an entire population. The diseases that are of particular interest are those with complex etiology, involving interaction of multiple genes and/or environmental factors. Such diseases (e.g., cardiovascular diseases, psychiatric illnesses, etc.) are common in the human population. Research towards the discovery of the genetic basis of these complex diseases is of major public health importance.

A tool for analysis of candidate genes and pathways involved in complex diseases: A common goal of human geneticists and genetic epidemiologists today is to dissect the molecular and environmental basis of common, heritable disorders. Linkage studies that have facilitated the identification of numerous single-gene, Mendelian disorders have had little success in study of diseases that arise from the interaction of multiple genes and the environment: that is, "complex" diseases. Alternatively, association studies have become a popular exploratory approach for genetic analysis of complex disease. In an effort to detect disease association, a number of single nucleotide polymorphism (SNP) markers, within a set of candidate genes deemed to be functionally significant, are genotyped in a sample of cases and controls. We present a software tool for such association analyses using a series of likelihood-based statistical tests, allowing for accurate p-value determination in the presence of complications such as massive multiple testing, conditional testing, and small sample sizes.

We adopted a likelihood-based approach because it allows for straightforward nesting of various hypotheses. The likelihood models evaluated include, (1) Hardy-Weinberg disequilibrium (HWD) in individual sample of cases and controls, (2) allelic association with disease (with and without HWD), and (3) genotypic association with disease [Emahazion et al., Trends Genet. 2001 Jul;17(7):407-13]. These likelihood models may also be extended to incorporate the effects of potential covariates, such as age, sex, environment, or known allelic/genotypic risk factor(s) at another locus. This is done by conditioning the likelihoods in both cases and controls on their status for the covariate and then performing the same battery of tests as before.

We also address the situation where the sample size is too small to apply the asymptotic theory for likelihood ratio tests in determining the significance of individual tests. To this end, we estimate empirical significance levels by randomizing the phenotype, and re-evaluating each statistic thousands of times to estimate statistical significance. The P-value for each statistic is defined as the proportion of randomized replicates in which the relevant statistic exceeds the value observed in the actual dataset. With the application of this randomization test, we can (1) determine the point-wise significance of each test per marker locus, (2) correct for multiple tests performed on each marker locus, (3) correct for multiple markers being tested per gene that results in inflated number of comparisons, and (4) test for "whole-pathway" effects by jointly considering all markers in a pathway.

Haghighi F., Terwilliger JD. A Tool For Analysis Of Candidate Genes And Pathways Involved In Complex Diseases. Cold Spring Harbor Laboratory, Genome Sequencing Conference.2002.

A bias-ed assessment of the use of SNPs in human complex traits: Although many biotechnological advancements have been made in the past decade, there has been very limited success in unraveling the genetic component of complex traits. Heavily

invested research has been initiated based on etiological models of unrealistic simplicity and conducted under poor experimental designs, on data sets of insufficient size, leading to an overestimation of the effect sizes of genetic variants and the quantity and quality of linkage disequilibrium (LD). Arguments about whether families or unrelated individuals provide more power for gene mapping have been erroneously debated as issues of whether linkage or LD are more detectable sorts of correlation. Although the latter issue may be subject to debate, there is no doubt that family-based analysis is more powerful for detecting linkage and/or LD. If the recent advances in biotechnology are to be exploited effectively, vastly improved study designs will be imperative, as the reasons for the lack of success to date have much more to do with biology than technology, an issue that has become increasingly clear with the findings of the past years.

Terwilliger JD., Haghighi F., Hiekkalinna TS., Göring HH. A 'bias'-ed assessment of the use of SNPs in human complex traits. Cur Op Genes and Development 12:726-734 (2002).

Evidence for a susceptibility locus for panic disorder near the catechol-O-methyltransferase gene on chromosome 22: BACKGROUND: A well-characterized single nucleotide polymorphism (472G/A-Val/Met-SNP8) in the coding sequence of the catechol-O-methyltransferase (COMT) gene leads to a three- to fourfold difference in enzymatic activity and clinical and animal studies suggest a role in anxiety states like panic disorder. METHODS: Subjects from 70 panic disorder pedigrees, and 83 "triads", were genotyped at seven single nucleotide polymorphisms (SNPs), polymorphic microsatellites in the first intron of COMT and approximately 339kb upstream of COMT (D22S944) and analyzed for genetic association and linkage. RESULTS: Linkage analysis showed elevated LOD scores for 472G/A (SNP 8), silent exon 3 substitution (186C/T-SNP 5), and the marker D22S944 (2.88, 2.62, and 2.93, respectively), using a variety of diagnostic and genetic models. Association tests were not significant for the SNPs, but were highly significant for D22S944 ($p = .0001-.0003$). One three-marker haplotype formed from the above three polymorphisms was significantly associated with panic disorder ($p = .0001$), as was the "global" p value for this combination ($p = .005$). In addition, numerous haplotypes with combinations of D22S944 and COMT SNPs were found to be significantly associated with panic disorder. CONCLUSIONS: Our findings provide strong evidence for a susceptibility locus for panic disorder either within the COMT gene or in a nearby region of chromosome 22.

Hamilton SP, Slager SL, Heiman GA, Deng Z, Haghighi F, Klein DF, Hodge SE, Weissman MM, Fyer AJ, Knowles JA. Evidence for a susceptibility locus for panic disorder near the catechol-O-methyltransferase gene on chromosome 22. Biol Psychiatry. 1;51(7):591-601 (2002).

Likelihood formulation of parent-of-origin effects on segregation analysis, including ascertainment: We developed a likelihood-based method for testing for parent-of-origin effect in complex diseases. The likelihood formulations model parent-of-origin effect and allow for incorporation of ascertainment, as well as differential male and female ascertainment probabilities. The results based on simulated data indicated that the

estimates of parental effect (either maternal or paternal) were biased when ascertainment was ignored or when the wrong ascertainment model was used. The exception was single ascertainment, in which we proved that ignoring ascertainment does not bias the estimation of parental effect, in a simple parent-of-origin model. These results underscore the importance of considering ascertainment models when testing for parent-of-origin effect in complex diseases.

Haghighi F., Hodge SE., Likelihood Formulation of Parent-of-Origin Effect on Segregation Analysis, Including Ascertainment. Am J Hum Genet. 70(1):142-56 (2002).

Stoppage: an issue for segregation analysis: Segregation analysis assumes that the observed family-size distribution (FSD), i.e., distribution of number of offspring among nuclear families, is independent of the segregation ratio p . However, for certain serious diseases with early onset and diagnosis (e.g., autism), parents may change their original desired family size, based on having one or more affected children, thus violating that assumption. Here we investigate "stoppage," the situation in which such parents have fewer children than originally planned. Following Brookfield et al. [J Med Genet 25:181-185, 1988], we define a stoppage probability d that after the birth of an affected child, parents will stop having children and thus not reach their original desired family size. We first derive the full correct likelihood for a simple segregation analysis as a function of p , d , and the ascertainment probability π_i . We show that p can be estimated from this likelihood if the FSD is known. Then, we show that under "random" ascertainment, the presence of stoppage does not bias estimates of p . However, for other ascertainment schemes, we show that is not the case. We use a simulation study to assess the magnitude of bias, and we demonstrate that ignoring the effect of stoppage can seriously bias the estimates of p when the FSD is ignored. In conclusion, stoppage, a realistic scenario for some complex diseases, can represent a serious and potentially intractable problem for segregation analysis.

Slager SL., Foroud T., Haghighi F., Spence MA., Hodge SE. Stoppage: An issue for segregation analysis. Genet Epidemiol 20(3):328-39 (2001).

An alternative method for computational detection of CpG islands: CpG islands are short segments of unmethylated DNA sequences, usually GC-rich DNA, that are dispersed throughout the comparatively GC-poor genome. CpG islands, generally .5 to 2KB in length, have approximately 60-70% G+C content and are highly conserved throughout evolution. It is generally believed that mammalian chromosomes are organized into domains with characteristic CpG island density, where the distribution of the islands is correlated with the 5 prime ends of genes such that the islands contain both the promoter and transcription unit (Kundu and Rao, J. Biochem., 1999, Cross et al., Mamm. Genome., 2000).

In the standard computational approach for detecting CpG islands, islands are defined as regions no longer than 200bp in length, with a moving average of $\%(C+G) > 50\%$ and ratio of observed versus expected CpG (O/E) > 0.6 , as

proposed by Gardiner-Garden and Frommer (J. Mol. Biol. 1987) and Larsen *et al* (Genomics, 1992). The identification of CpG islands involves sliding a window across the chromosome, where the window size and sliding distance is specified in an *ad hoc* fashion.

We propose an alternative method based on recursive segmentation, for delineating the location of CpG islands. We first preprocess the data, converting the DNA sequence into a binary sequence with 1 corresponding to an observed base in a CpG dinucleotide and 0 otherwise. In this way, CpG islands are characterized as regions with a high density of bases coded as 1. The recursive segmentation algorithm can identify subsequences with homogeneous 0 or 1 base composition, corresponding to potential boundaries/partitions of CpG islands. The segmentation algorithm requires a stopping criterion used in determining the maximal partition. The stoppage threshold used in our analyses is based on the Bayesian information criterion, which is a function of sequence length and number of model parameters before and after segmentation (Li, Proc. RECOMB01, 2001). The use of a defined stopping threshold is similar to the above method's use of a defined threshold for CpG dinucleotides [i.e., (O/E) > 0.6].

We have implemented the recursive segmentation algorithm, which allows for processing of the entire chromosomal sequence in an efficient manner. Our preliminary results from analysis of chromosome 22 sequence are promising and reveal a further direction for improvement, in particular, increasing the stringency of the stopping threshold, since the current threshold proved to be too lax and yielded a large number of very small CpG islands. In this study we report the predicted genome-wide CpG islands and compare the results with the standard approach as well as determine the correlation of these CpG islands with annotated genes and gene features in the human genome.

Haghighi F., Grundy NB., Li W. An Alternative Method for Computational Detection of CpG Islands. Cold Spring Harbor Laboratory, Genome Sequencing Conference. 2001.

Li W., Bernaola-Galvan P., Haghighi F., Grosse I. Applications of recursive segmentation to the analysis of DNA sequences. Comput Chem. 26(5):491-510 (2002).

Mapping the methylation landscape of the human genome: Intact genomic methylation patterns are necessary for the proper functioning of the genome, yet very little is known of their actual form. We have developed novel experimental and computational methods to map the methylation landscape of the genome. This involved fractionation of the genome into methylated and unmethylated compartments and preparation of plasmid libraries from these compartments. High throughput sequencing of large numbers of these clones provided us with sets of methylated and unmethylated domains from normal human brain. To analyze the large amount of sequence data thus generated, we have developed software that automatically assembles and identifies the inferred methylated and unmethylated regions within the human genome. Given a collection of paired sequence reads as input, the software searches the genome and locates all exact or near-exact matches of each given sequence read, assembles the complete corresponding library fragments by extracting the intervening sequences, and determines the methylation status of the

region based upon the presence or absence of the restriction enzyme sites by which the library was constructed. The methylated and unmethylated regions are further characterized by examining the local and global genomic features within these regions. For the local analyses, the individual sequences are evaluated using a range of sequence-based metrics (e.g., CpG dinucleotide frequencies and G+C content) and further assessed using a graphic interface to existing genomic tracks in the Genome Browser (<http://genome.ucsc.edu/>). For the global analysis, we measure the association between genomic features (e.g., CpG islands, transposons and other repeated sequences, exons, introns) and methylation state. New graphical methods have been developed to display the relationship of methylation status to content of a large number of diverse sequence features. Together these analyses have revealed striking patterns in the clustering of CpG rich sequences in the unmethylated regions, where Alu elements are found at the boundaries of kilobase-sized unmethylated domains. As expected, most methylated domains are rich in transgenes while unmethylated domains contain few transposons other than those severely eroded by mutation. The goal of this study is a set of criteria whereby the methylation status of a given sequence can be predicted with high confidence.

Haghighi F., Rolins RA., Bestor TH., Mapping the Methylation Landscape of the Human Genome, Cold Spring Harbor Laboratory, Genome Informatics Conference.2003.