THE ROLE OF STATISTICS IN PHARMACOGENOMICS

Sandra Close Kirkwood*

ABSTRACT

The sequence of the human genome released earlier this year is a significant scientific achievement. The goal of pharmacogenomics is the application of genetic information and technology to develop better therapeutics or to guide the use of pharmaceuticals in the treatment or prevention of disease. Statistical theory and probability will play an expanded role in interpreting genetic information through the development of new analytical methodology and the novel application of traditional statistical theory. Examples in gene mapping and microarray expression analysis will be used to broadly illustrate the essential role of statistical theory in pharmacogenomics research. Specific gene mapping methodologies discussed include linkage analysis, linkage disequilibrium studies, and haplotype analysis. Application of statistical theory to gene chip experiments to obtain high-quality data including experimental design, minimizing variability, and well-controlled verification strategies and applications to identify gene expression differences between experimental groups will be reviewed. The combination of statistical applications and genomic technologies is key to understanding the genetic differences that identify patients susceptible to disease, stratify patients by clinical outcome, indicate treatment response, or predict adverse event occurrences.

1. Introduction

Recently two independent groups released the first draft of the human genome (Lander et al., 2001; no author, 2001) resulting in a slate of editorial comments about how this knowledge will revolutionize medicine and drug development (Collins and McKusick, 2001). Sequencing the human genome is a significant scientific achievement; the question now is whether we can capitalize on the information to cross a new frontier in medicine. This genetic information should allow us to assess an individual's risk not only for rare inherited diseases but also for common health problems including cancer, hypertension, arthritis, heart disease, and mental illness. The application of this genetic information in the pharmaceutical industry should result in new therapies designed not only to treat symptoms but prevent disease and allow the prediction of an individual's response and likelihood of developing side effects to a particular treatment.

The sequencing of the genome and advances in molecular genetic technologies have enabled scientists to go beyond the study of a specific gene or pathway known to be involved in disease or drug metabolism to discovering new pathologic pathways from knowledge of the structure and function of the genes. Functional genomics and proteomics will allow us to gain insight into the basic mechanisms of cell function. These advances in molecular biology have resulted in a paradigm shift in the questions biologists in the laboratory are asking statisticians. Historically, a typical question to a statistician might have been

*Eli Lilly and Company Lilly Corporate Center DC 0444, Indianapolis, Indiana 46285, USA

Key words: Analysis; Associations; Genomics; Linkage; Microarray
in reference to one sequence or regarding the biological expression of one gene. Now the
question might involve the interpretation of thousands of sequences or thousands of genes.
Experiments generate information that awaits interpretation, rather than being designed to
answer one specific question or hypothesis. These experiments require efficient information
management, the creation of databases to store and retrieve the data, large-scale sequenc-
ing capabilities, functional predictions for the sequences, and clustering algorithms to group
the gene expression. Initially the majority of the analysis of genomic data was completed
by bioinformatics. Bioinformatics will continue to fulfill a vital role in the genomic revo-
lution working closely in conjunction with statisticians. Statistical theory and probability
will play an expanded role in interpreting the information gleaned from the genomic revo-
lution through the development of new analytical methodology and the novel application of
traditional statistical theory.

2. Pharmacogenomics

Pharmacogenomics is the application of genetic information and technology to develop
better therapeutics or to guide the use of pharmaceuticals in the treatment or prevention
of disease. Pharmacogenomics has applicability across the drug development pipeline from
assisting in the identification of better targets (Bumol and Watanabe, 2001), predicting tox-
icty or metabolism response, developing biomarkers, and designing optimal clinical studies.
Pharmaceuticals will be developed to block or stimulate a particular pathway known to be
involved in disease pathogenesis. Examples of targeted medications include Herceptin®,
a humanized antibody against her-2 neu developed for the treatment of her-2 neu over-
expressing breast cancer (Cobleigh et al., 1998) and Glivec®, designed to treat chronic
myelogenous leukemia by blocking the activity of the bcr-abl kinase (Druker, 2000).

In addition to designing targeted therapeutics, genetic information will be used to
predict response and toxicity ultimately improving outcomes and decreasing risk. Genetic
variability is recognized as not only one of the main causes of disease but of variable drug
response. Identification of genes leading to susceptibility for disease such as apolipoprotein
E4 and risk for Alzheimer’s disease (Roses and Saunders, 1994), in addition to providing
new targets for pharmaceuticals, may potentially affect response to therapeutics. Other
genes of interest that definitely affect drug response include the pharmacokinetic genes such
as the cytochrome P450 enzymes such as CYP2D6. CYP2D6, responsible for metabolizing
approximately 25% of the current cadre of commercially available drugs, has genetic variants
that result in absent, decreased, or enhanced enzyme activity leading to the poor and ultra-
rapid metabolizer phenotypes (Meyer and Zanger, 1997). Genetic variants in the drug
targets themselves can also affect drug response such as the polymorphism in the regulatory
region of the 5-lipoxygenase gene that affects response to asthma medication (Drazen, 1999).
Identification of variants such as these has the potential to change the prescribing practices
of physicians allowing a more individualized approach, tailored to an individual’s genetic
profile (Roses, 2000).

Statistics contributes significantly to the current pharmacogenomics tools such as com-
putation of statistical significance for sequence alignment and data base searching, likelihood estimation for phylogeny construction, and using Markov Models to describe biological structures. However, the polygenic and multifactorial etiology of common diseases, drug response, and the development of side effects makes harnessing genetic knowledge into treatments difficult. For example, the relationship between different genes and between genes and environmental factors are poorly understood. Probability theory and statistics plays
an increasingly important role in pharmacogenomic analyses by identifying the relevant genetic factors, understanding their relationship to other genes and environmental factors, and validating the association between genetic markers and drug response or side effect development.

We are at the beginning of a long pursuit whose ultimate goal is better-targeted pharmaceuticals and the improvement of patient care. Statistics will play a major role in differentiating the reality from the hype. The purpose of this manuscript is not to give a comprehensive review of the methodology or applications of statistics in pharmacogenomics but to use examples that broadly illustrate the essential role of statistical theory in pharmacogenomic research.

3. Gene mapping

Since R.A. Fisher, statistical methods have been integral when the questions involve looking at the relationship between a phenotype or trait and genetic diversity to locate a disease gene. Since 99.9% of the genetic code is the same from person to person (Collins, 1997), the focus of pharmacogenomics is on the 0.1% that varies between individuals. Polymorphisms are genetic markers of diversity present in at least 1% of the population defined by the type of mutation resulting in the polymorphism. A single nucleotide polymorphism (SNP) arises from a single base pair mutation substituting one nucleotide for another. Other polymorphisms result from the insertion or deletion of a section of deoxyribonucleic acid (DNA), frequently a repeated sequence of bases ranging in size from two to several hundreds of base pairs. These repeat polymorphisms are variable and thus are highly polymorphic. The most common are microsatellite markers and their variability makes them useful for population genetic studies since the probability that two individuals from different populations will have the same number of repeats is remote. Humans have 2 copies of every gene, one paternal and one maternal. The term allele will be used throughout this manuscript to refer to each individual copy of a gene.

3.1. Linkage analysis

Traditional linkage analysis or meiotic mapping capitalizes on the recombination that occurs during meiosis I to identify chromosomal regions harboring genes related to drug response, side effects, or disease susceptibility. Linkage analysis methods assess the transmission and co-segregation of alleles at marker loci, or landmark places on the genome, with phenotypic alleles assumed to be carried by affected members of a pedigree. Tracing the consistency of co-segregation of the marker locus and phenotype from generation to generation allows localization of putative genes to a chromosomal region. Through the efforts of multiple academic and industry initiatives many genetic markers have been catalogued to use in this pursuit. Over the past several years polymorphic microsatellite markers have been used in gene mapping efforts. Recently, SNPs have emerged as an alternative marker. SNPs are relatively easy to genotype due to their simple structure as base changes; they are frequent and located throughout the genome in exons, introns, intergenic regions, promoters, enhancers, and are generally less mutable. These characteristics give them utility for trait or disease gene discovery.

A variety of study design and statistical methodologies are used for linkage analysis (Rao and Province, 2001). Parametric linkage analysis involves specifying a model of inheritance and testing the null hypothesis that the phenotypic gene is not linked to a given genetic marker versus the alternative that a specific chromosomal region contains the putative gene
and it is close to a known genetic marker. A likelihood ratio is then used to evaluate the strength of evidence of linkage relative to evidence for no linkage. The multifactorial nature of common diseases makes defining the pattern of inheritance difficult; thus, nonparametric linkage models, which do not require specifying the inheritance, are frequently applied to complex diseases. Allele sharing methods determine whether related individuals with the phenotype of interest, such as responders or non-responders to a certain drug, inherit the same genomic region more often than expected if the phenotype and genomic region segregated independently. This method often uses sibling pairs. Two methods frequently used for qualitative traits are to compare the observed and expected numbers of sibling pairs sharing zero, one or two alleles using a simple chi-square with two degrees of freedom and inferring probabilities of the shared alleles under constraints imposed by linkage and then computing a statistic such as the maximum likelihood of odds score. For quantitative traits, sibling pair linkage analysis involves linear regression of the squared differences in trait values within sibling pairs on the proportion of alleles shared. Individuals more similar are expected to share more alleles near the putative gene, a regression line with a negative slope. Variance component-based linkage analysis is another statistical tool to evaluate quantitative trait loci influencing complex phenotypes. Linkage methods may involve simultaneous consideration of multiple markers extracting the maximum amount of inheritance information.

For the past several years, linkage analysis was followed by physical mapping or positional cloning of a specific gene thought to play a role. This method was extremely rewarding with greater than 1000 disease genes located including the cystic fibrosis, Huntington disease, and Duchenne muscular dystrophy genes. However, in many instances, linkage analysis has not proven powerful enough to detect genes with small to moderate effect as is the case with multifactorial phenotypes. Thus, an important aspect of study design is evaluating the power to detect a true linkage. The ability to detect linkage is dependent on a number of factors including the number of identifiable meiotic events, which is a function of the number, size and structure of the families, the quantity and informativeness of the markers, and the variance attributed to the putative gene. Simulation studies estimating these parameters are required prior to sample size determination or following collection to estimate the power of a sample.

3.2. Linkage disequilibrium

An alternative method to identify a chromosomal region harboring a putative allele, used independently or in conjunction with linkage analysis, is linkage disequilibrium or association studies. Association studies often compare the frequencies of alleles between unrelated cases and controls. An allele is considered to be associated with a trait if the allele occurs at a higher frequency in the case group relative to controls. The association of a disease phenotype and a genetic polymorphism in a population may be due to: 1) an allele at the locus contributing to the trait, 2) an allele at the locus in disequilibrium with a true trait allele nearby, or 3) a spurious association due to the sample containing subpopulations or mixtures of individuals with different genetic backgrounds. Although association between a genetic variant and disease does not establish causality, association analyses may be more powerful and sensitive to genes of modest to small effect.

Genome wide association studies using SNPs have been proposed due to the abundance of SNPs (Kruglyak, 1999). Linkage disequilibrium is a function of time since the mutational event and the distance from the marker to the putative allele; therefore, the size of the genomic region harboring alleles that co-segregate may be very small. Knowledge of how far linkage disequilibrium extends and the strength of the disequilibrium is crucial for
The Role of Statistics in Pharmacogenomics

disequilibrium-based mapping studies and can help guide the sample size and the number and density of markers needed for a study (Kaplan and Morris, 2001). A dense map of markers increases the resolution and power of the study. However, complexities arise due to the high rate of genetic recombination and the influence of many genes on common traits making it difficult to identify associated SNPs (Schork et al., 2000). Crucial are analytical techniques and study designs that can be used to assess evidence of co-segregation and linkage disequilibrium. In addition, the testing of these tremendous numbers of polymorphisms raises statistical issues related to false positive rates and the level of significance.

Due to advances in molecular biology, available candidate genes, and the increased number of available genetic markers (Schidanandam et al., 2001), association studies using candidate genes are becoming a viable option for pharmacogenomic studies. A variety of sources may provide candidate genes to test for linkage disequilibrium with a phenotype of interest. If linkage analysis has provided us with conclusive results as to the chromosomal region of a morbid locus, the genes present in this region can be screened for mutations in patients' genomes and tested for association with the phenotype. Other genes may be identified purely on biological hypothesis and known function of the gene, from the literature, from experiments using model organisms, from in silico methods to uncover relevant genes from sequence data, or from the application of new technology such as gene chip experiments comparing the gene expression patterns in treated and non-treated or diseased and non-diseased tissue (Niculescu et al., 2000).

Statistics often applied in association studies utilizing a case-control design include Chi-square, analysis of variance, and conditional logistic regression. Family based association methods have been proposed to help decrease the rate of false positives as a result of population substructure (Bull et al., 2001). A common family based association test is the transmission disequilibrium test that exploits the transmission of alleles from parents to children utilizing McNemar’s test to test for significance. Statistical theory has been integral to the development of additional methods to reduce the false positive rate in disequilibrium studies including adjusting for population substructure using multi-locus genotype data to infer population structure and ancestry and make the correct adjustments to the significance level (Satten et al., 2001; Reich and Goldstein, 2001; Devlin and Roeder, 2000; Pritchard, 1999). As with linkage studies, statistical methods are necessary to evaluate of the power of the collection scheme and sample size, and to compare the various proposed methods and experimental designs for disequilibrium studies (Abecasis et al., 2001; Schork et al., 2000; Page and Amos, 1999; Tu and Whittemore, 1999; Allison, 1998).

3.3. Haplotyping

Linkage disequilibrium mapping can be facilitated through haplotype analysis. A haplotype is several neighboring SNPs surrounding an ancestral mutation that are transmitted to descendants. Alleles show distinctive patterns of linkage disequilibrium due to characteristics of their evolution creating haplotypic diversity with direct application to both linkage and association studies. Haplotypes are more powerful than focusing on a single locus; however, difficulties lie in haplotype determination. Due the humans being diploid organisms, molecular haplotyping (direct determination) is only possible through the laborious and expensive process of biologically sequencing the paternal and maternal chromosomes independently. If a molecular haplotype is unknown then the phase of the chromosomes must be inferred using algorithms based on the analysis of the SNPs in individuals and information on haplotype frequency in the population.

Multiple statistical methodologies such as Clark’s algorithm (Clark, 1990), phyloge-
netic analysis (Lam et al., 2000), applications of the Expectation-Maximization algorithm (Clayton, 1999; Long et al., 1995; Excoffier and Slatkin, 1995), Markov chain (Kuhner and Felsenstein, 2000), and phase reconstruction methods (Stephens et al., 2001) have been utilized to estimate haplotypes. The best available algorithms only correctly assign 80-90% of haplotypes with the percentage decreasing dramatically with more complex haplotypes. This level of accuracy is not acceptable in many instances in pharmacogenomics and better algorithms to capture all available information regarding SNP frequency are needed. In order for haplotyping to provide more power than evaluating single SNPs the increase in linkage disequilibrium must be large enough to compensate for the increase in degrees of freedom of the chi-square test (Kaplan and Morris, 2001). In addition, population considerations such as the frequency of the alleles at the disease mutation and in the population at large affect the power. More sophisticated haplotype methodology considering these factors is needed (Horikawa et al., 2000). In order to achieve the most accurate estimate of haplotypes, statistical methods can be used in conjunction with experimental methods.

Recent methodological developments to address some of the issues with traditional linkage and association studies include applications of neural networks to gene localization, recursive-partitioning models to address heterogeneity, and methodology for meta-analysis. A number of approaches should be used to identify genes of interest and the combination of information regarding causative SNPs with good phenotyping and rigorous statistical methodology will capitalize on the available genetic information.

4. New technology

In addition to contributing significantly to the development of gene mapping methodology, statistical input is essential to the successful integration of new technology into pharmaceutical research. Molecular biology has been revolutionized with the development of novel sequence and expression analysis tools. These tools generate massive amounts of data and statistical principles must be incorporated in the management and analysis of these data sets. Furthermore, statistical theory is essential in evaluating the reliability and validity of any new technology. Gene expression microarrays or gene chips are the most prominent technological advance to date; therefore, I will use expression data to illustrate statistics vital role in the analysis of any data created by implementation of this technology.

4.1. Gene chips

Nearly all the cells in the human body carry exactly the same set of genes. Many of which are turned on in the majority of cells; however, a unique subset of genes is expressed to give cells their distinct characteristics which make neuronal cells neuronal and muscle cells muscle, etc. Gene chips can be used to identify the cellular pattern of gene expression, which is essentially a genetic fingerprint of that cell type. A gene chip is similar to a microchip but it is coated with DNA rather than electronic circuitry. Thousands of strands of reference DNA are synthesized on the chip with photolithography, ink-jet spray, or pin spotting. The RNA to be tested is labeled with fluorescent dyes and placed on the chip where binds to the complementary DNA. The luminescent pattern is then measured and fed into a computer for analysis. There are many different microarray technologies, two widely used within the research community include Affymetrix oligonucleotide microarrays produced by a photolithographic process and cDNA microarrays utilizing polymerase chain reaction amplified fragments of DNA onto a solid support. These two technologies will be used as examples when discussing statistical issues and will be referred to as oligonucleotide
arrays and cDNA arrays.

The data produced from a gene chip experiment is a great example of the paradigm shift mentioned previously. The expression of literally thousands of genes can be investigated using the chip. The current challenge is to extract useful and reliable information from these large data sets. Technological problems and biological variation can make it difficult to distinguish signal from noise. The noise level is dependent on the tissue or cell being used, the experimental design, and the specifics of the technology being used. Statistical theory has a proven track record in validating assays and reducing experimental noise, why should microarray experiments be treated differently? Variability may be introduced throughout the experimental process from a variety of sources both common across different methodologies and unique to the specific technology being utilized. Variation in sample characteristics and preparation including RNA isolation, clean up, and labeling affect the quality of the starting material. The quality of data from cDNA experiments can be improved by replication of spots, reversing the labeling process, and careful experimental design regarding placement of the spots. In oligonucleotide chip experiments factors such affecting chip to chip variability must be considered. Considering the amount variability is critical to avoid experiments where the variability between groups or samples is less than the variation between replicates of the same sample. Technical problems are common with any new technology including microarray experiments; however, early recognition of these will save time and resources. In addition to assessing the variability of the system and the reliability of a single hybridization, assessing the reliability of the biological effect raises critical issues related to the design of the number and types of replicates and numbers of animals or individuals to include in each group. Application of statistical theory to experimental design including issues of quality control, minimizing variability, and well-controlled verification strategies are essential for obtaining high-quality meaningful data.

Application of this information includes identification of genes related to disease process, drug response, or development of side effects. Current analytical approaches include cluster analysis. A variety of methodologies have been employed including hierarchical clustering, K means clustering, self-organizing maps, and principal component analysis. Cluster analysis whose primary objective is to group genes by comparable patterns of variation is useful for reducing the complexity of large data sets and for identifying predominant patterns within the data. Group analysis (Mirmics, 2000, 2001) clustering genes together by structure or function, and then using measures of central tendency to compare the overall gene expression distribution of the groups compared to the overall gene expression has been employed to detect putative function-related expression changes. However, the utility of these is limited by the extent of our current knowledge regarding gene structure and function.

Specific pharmacogenomic applications include using this technology to identify differences between experimental groups leading to new therapeutic targets and the identification of genetic markers of drug efficacy and side effect development. To different gene expression between experimental groups, traditional methodology such as analysis of variance with adjustment for multiple testing is often employed. Additional methodology is continuously being proposed in the literature. Examples include the single pulse method recently published by Zhao (Zhao et al., 2001), the method by Tusher to evaluate significance by assigning a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements with permutations to minimize confounding effects (Tusher et al., 2001), and statistical modeling using heterogeneity parameters to reduce the impact of noise in the data sets and provide methods to reduce the number of genes to be ana-
lyzed. Currently there is much debate in the literature regarding the “appropriate” way to analyze array data and a shift appears to be occurring towards a more statistical approach. However, it is clear that data analysis methods require substantial improvements and the successful application of this technology is going to be dependent on statistics working in conjunction with bioinformatics to garner the most information from the data.

These same principles for data analysis and quality control have application to any new technology introduced into a pharmacogenomics laboratory. In the future microarray sensitivity, complexity and availability will increase, protein and tissue arrays will become more reliable and available, sequence based microarrays will correlate expression changes with informative polymorphisms, pharmacokinetic, or drug dose and blood concentration and pharmacodynamics or measurements of drug response including genetic polymorphisms will be linked, and data analysis tools will allow data to be analyzed across platforms. Mathematical models and statistical theory will be necessary to thoroughly understand the data resulting from these new technologies will be vital to understanding dose response, side effects, designing more targeted therapeutics and for earlier go or no-go decisions.

5. Conclusion

A combination of statistical applications and genomic technologies is the future of medical and pharmaceutical research. In the near future genomics may enable differentiation from the competition and the development of targeted therapeutics. In the midterm relevant genetic variations associated with clinical response, improved efficacy and decreased toxicity, to a pharmaceutical will be identified leading to improved labeling. Based on an individual’s polymorphism profile at relevant genes physicians will prescribe the medication best suited for a particular individual. In addition, increased genetic information may allow for better prediction of prognosis. The long-term goal is targeted pharmaceuticals based on known pathophysiology and the relationship of disease to genetic polymorphisms leading to preventative medicine. However, prior to achieving any of these “promises” of the genetic revolution, technology must be carefully applied and understood and the relevant genes identified. The application of statistics in genomics is vital to both of these endeavors.

The challenges in statistical analysis include issues related to study design, reproducibility of data, scalability of methods, missing data and power calculations. Communication channels to statistics need to be open. In order to capitalize on the data resulting from the application of these new technologies, molecular biologists and bioinformaticians must interact closely with statisticians. Statistics should be an integral essential part of the planning, implementation, and analysis of future experiments. As a result of the technological advances in molecular biology techniques, a wealth of data is available for analysis. The key will be in the analysis of that data to reveal the many secrets, which will ultimately impact drug discovery and development. We are at the beginning of a long pursuit whose ultimate goal is the improvement of patient care. The understanding of the genetics that identify patients susceptible to disease, stratify patients by clinical outcome, indicate treatment response, or predict adverse event occurrences is in its infancy. The potential is enormous but without careful application of statistical theory throughout the developmental process this potential will be difficult if not impossible to achieve.
REFERENCES


