The Need for a Systematic Approach to Complex Pathways in Molecular Epidemiology

Duncan C. Thomas
Keck School of Medicine, University of Southern California, Los Angeles, California

Thinking about biochemical pathways has become an increasingly important part of molecular epidemiology. The field is rapidly moving from evaluation of single candidate genes, one at a time, to consideration of entire pathways comprising perhaps dozens of genes and their environmental substrates, even multiple pathways that link up or compete in complex networks. Even in its simplest rendering, for example, dietary folate seems to be a protective factor for colorectal cancer and is involved in at least two distinct pathways, with relative activities regulated by the methylentetrahydrofolate reductase protein (among other factors): one involves DNA methylation; the other involves disruption of pyrimidine synthesis, leading to increased DNA damage and repair (1, 2). Many proteins critical in folate metabolism are coded for by genes with known polymorphisms. Alcohol and vitamins B6 and B12 also play a role in the folate pathway, and of course other pathways involving metabolism of heterocyclic amines, polycyclic aromatic hydrocarbons, bile acids, and nonsteroidal anti-inflammatory drugs might compete or interact with the folate pathways in ways we can only speculate on (3). (Whereas epidemiologic evidence for a role of heterocyclic amines in colorectal cancer is weak, its effect may be diluted because we have not considered full pathways and because the proxies for heterocyclic amines are more difficult to measure than the proxies for polycyclic aromatic hydrocarbons, leading to a greater degree of nondifferential misclassification of heterocyclic amine exposure than polycyclic aromatic hydrocarbon exposure.) To further complicate matters, there is evidence that folate may protect against early precancerous lesions but increase the risk of cancer in those with preexisting lesions (4).

Beyond pathways guided by prior physiologic knowledge and genetic variants suggested to be functionally significant, the advent of new genomic tools now makes it possible to characterize the full spectrum of genetic variation within candidate genes using haplotype-based methods (5). The technology will also soon allow genome-wide searches for gene associations and interactions, adding powerful exploratory tools and a whole new level of complexity (6). Whereas pathway-driven research is an important step forward, thus far it has been used primarily to select promising candidates for genetic characterization and study of plausible gene-environment and gene-gene interactions. Analytic tools to use this wealth of data are still in their infancy. Large-scale case-control, cohort, and family-based studies are currently under way that will assess many polymorphisms in scores of genes, and in combination with environmental factors (7, 8). Still, most reports from such studies are limited to relative risk estimates from each factor considered one at a time or in pairwise combinations, using very traditional epidemiologic analysis tools. There is an obvious reason for this: even with the largest studies, statistical power for testing even two-way interactions is often limited and finer stratification by three or more factors rapidly leads to inadequate sample sizes and unstable risk estimates. Even main effects of candidate gene associations have proven notoriously difficult to replicate (9) and reported interactions even harder (10, 11). A recent editorial in this journal (12) lays out criteria aimed at “improving the environment for publication of association studies,” including serious consideration of biological plausibility and pathways. The difficulties of multiple comparisons that will arise in the next generation of genome-wide association scans are daunting (6, 13, 14). Demands for very small P values or low false discovery rates will erode power for testing interactions further. Whereas there will always be a place for exploratory data analysis techniques (data mining, neural nets, classification and regression trees, multidimensional reduction, clustering, etc.; refs. 15-21), our hope is that incorporation of prior biological knowledge about pathways will enhance the ability to detect real causal effects. To fully realize this hope, however, we need better analysis tools that will fully exploit this knowledge and epidemiologic study designs that will provide the necessary data.

To avoid the “curse of dimensionality,” some kind of structure is needed for any complex statistical model to be estimable. Simply eye-balling a large table of unconstrained relative risks for various cross-tabulations of genes and exposures is unlikely to be very rewarding, although such tables can be a useful starting point for exploratory analyses and for communicating the results of a more sophisticated model. Standard multivariate modeling approaches, such as logistic regression, are one way of putting structure on a model, so as to focus attention on main effects adjusted for each other and to test interactions in a natural hierarchical sequence [e.g., all (or a subset of “interesting”) two-way interactions, followed by three-way interactions that involve combinations of significant main effects and two-way interactions; ref. 22]. However, inevitably an element of subjective choice between competing models creeps in, unless one adopts some purely mechanical stepwise procedure, an approach that is well known to be unlikely to uncover the true model. Hence, the interpretation of the effects included in one’s “best” model and the variances of their estimated coefficients somehow needs to allow for this uncertainty about model choice (23). Hierarchical mixed models, which treat the regression coefficients as random effects with some common distribution (24), and Bayesian model averaging across the space of all possible models (25, 26) are two approaches that aim to account for this problem of model uncertainty. Another approach to putting structure on a complex system is to adopt a mechanistic model, such as physiologically based pharmacokinetic (PBPK) models for metabolic pathways (27, 28). These typically assume some form of Michaelis-Menten kinetics for each step of a pathway and use differential equations to predict the relationship between substrate concentrations and the genes that determine the various reaction rates. Such models are widely used in toxicology (29-31), but until recently, only limited attention has been given to allowance for interindividual variation in the kinetic variables and their dependence on genotypes. Although a promising approach for metabolic...
pathways, such as heterocyclic amine and polycyclic aromatic hydrocarbon metabolism, different types of models will doubtless be required to describe, say, DNA repair and cell cycle control pathways, and feedback loops pose formidable mathematical challenges.

Early applications and simulation studies of such approaches (hierarchical Bayes and PBPK models) indicate potentially serious problems of identifiability and lack of robustness to modeling assumptions if the only data available are from a traditional epidemiologic study with measurements of exposure, genes, and outcome, together with a priori knowledge about the topology of the network but not about the kinetic variables or intermediate metabolites themselves. To move forward, it is essential that more information be incorporated into the analysis of the available epidemiologic data. Biomarkers offer one particularly tempting source of additional data (32, 33). These could take a number of forms—markers of exposure (e.g., measures of heterocyclic amine content of various forms of cooked meats or validation studies of questionnaire-based dietary assessments using biomarkers of heterocyclic amines; ref. 34), circulating or excreted concentrations of intermediate metabolites, enzyme activity levels, DNA adducts, damage, methylation, or even preclinical markers of disease. Ideally, such measurements would be obtained longitudinally and would precede the onset of disease, requiring a cohort design. Studies of the familiality of such markers, to allow for unmeasured genetic or environmental determinants, or even linkage studies to localize other genes that influence the metabolic rates could be enlightening.

The cost and invasiveness of such studies will likely preclude obtaining such detailed information routinely on large cohorts or even on case-control samples, where the possibility of disease altering the pathway variables among the cases would be a problem. However, it would be possible to incorporate information from external validation studies (say, longitudinal observation of a small number of individuals in a metabolic lab), or better yet, use of multitissue sampling designs within large-scale epidemiologic studies (35, 36). Hierarchical and PBPK models are naturally designed to allow the incorporation of such data in the higher levels of the model. For example, unobserved person-specific metabolic rates in a PBPK model can be regressed on genotype, or relative risk variables in a hierarchical model can be regressed on functional assays or in silico predictions of genetic effects; see Hung et al. (37) for an example of the use of pathway indicators as prior covariates in a traditional epidemiologic model. Another recent article (38) provides an example of the use of a randomized crossover design to investigate a gene-environment interaction on an intermediate phenotype, in this case between GST design to investigate a gene-environment interaction on an

Computational algorithms such as PolyPhen (41), SIFT (42), and the Grantham scale (43) have been developed to characterize DNA sequence data in terms of evolutionary conservation and predicted effects on protein conformation (44); such predictions have been shown to be correlated with relative risks across a spectrum of epidemiologic associations (45-47). Information from animal models as well as gene expression, proteomic, and epigenetic studies could also be incorporated explicitly in epidemiologic modeling. For example, one might use data from expression, proteomics, or siRNAs to identify gene sets that may not be part of recognized pathways; this approach is illustrated by the recent use of yeast deletion consortium data in a genome-wide scan for genes affecting radiosensitivity (48).

When conceptualizing a study, we frequently think we know a lot about the structure of a pathway, but as we get deeper and deeper into the biological literature they invariably become more complex and uncertain. Whereas we would like to be able to exploit what we think we know from other sources in developing models that can be applied to the epidemiologic data at hand, we do not want the results to be foregone conclusions, driven more by our assumptions than by the data itself. Statistical methods such as Bayes factors (49) are available to quantify the incremental contribution of the data beyond the assumed structure of an analysis, but these need to be nested within a sufficiently broad class of models to allow for the possibility that our prior knowledge is wrong, without at the same time being so flexible as to completely defeat the purpose of trying to use our biological understanding to give some structure to our analyses. One of the aims of the field of systems biology is to “reverse engineer” a biological system to infer from observations the underlying structure, hardware, and software a cell uses to achieve some function (39, 50, 51). Techniques such as Bayesian network analysis have been useful for this purpose (52). This reverse engineering is precisely the aim of molecular epidemiology, at the opposite extreme of complexity—populations rather than cells—so we would do well to incorporate some of the thinking that systems biologists and other disciplines use for this purpose.

In summary, pathway-driven thinking is potentially a major step forward for the field of molecular epidemiology, by providing a unifying framework for the investigation of multiple genes and environmental factors that act in concert. This hope will only be realized, however, by advances in study design and statistical analysis that can integrate all these factors into a single model, incorporating prior knowledge from such disciplines as molecular genetics, biochemistry, toxicology, and systems biology. The rapid accumulation of knowledge from such fields into Omics databases and the potential to measure intermediate biomarkers provide exciting opportunities to support this development.

References

The Need for a Systematic Approach to Complex Pathways in Molecular Epidemiology

Duncan C. Thomas