Editorial

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 methylenetetrahydrofolate reductase protein (among other factors): one involves DNA methylation; the other involves disruption of pyrimadine 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

Cancer Epidemiol Biomarkers Prev 2005;14(3):557–9

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 multistage 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 hierarchical model for bladder cancer. Another study (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* genotypes and diesel exhaust particles on acute allergic responses, data which might later be incorporated into the analysis of a chronic disease endpoint using a hierarchical model.

In addition to actual measurements that could be made within a specific epidemiologic study, we need to get smarter about tapping into the wealth of data available in ''Omics'' databases (39) or not even catalogued but obtainable by close collaboration with colleagues in other disciplines. Several databases, such as the Kyoto Encyclopedia of Genes and Genomes (www.genome.jp/kegg/), provide a wealth of information about the structure of biochemical pathways, including genetic determinants and the available rate variables, whereas De Roos et al. (40) have appealed for an integrated ''Exposure-Gene-Disease'' database aimed specifically at the needs of molecular epidemiology and shown how such data could be exploited in the hierarchical modeling framework. Development of functional assays for specific polymorphisms is a labor-intensive process, but once in place could allow all polymorphisms in an interesting candidate gene to be characterized in ways that could have some bearing on their predicted risk (e.g., for a study of *ATM* variants, expression levels, kinase activity, cell cycle checkpoint activation, and colony survival of cells heterozygous for variants after exposure to ionizing radiation).

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 models. For example, one might use data from expression, proteomics, or siRNAs to identify genes 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

1. Ulrich CM, Robien K, McLeod HL. Cancer pharmacogenetics: polymorphisms, pathways and beyond. Nat Rev Cancer 2003;3:912–20.
2. Ulrich CM, Robien K, Sparks R. Pharmacogenetics and folate metabolism–a promising direction. Pharmacogenomics 2002;3:299–313.
3. Potter JD. Colorectal cancer: molecules and populations. J Natl Cancer Inst 1999;91:916–32.
4. Kim YI. Will mandatory folic acid fortification prevent or promote cancer? Am J Clin Nutr 2004;80:1123–8.
5. Gibbs RA, Belmont JW, Hardenbol P, et al. The International HapMap Project. Nature 2003;426:789–96.
6. Lin S, Chakravarti A, Cutler DJ. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat Genet 2004;36:1181–8.
7. Clayton DG, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001;358:1357–60.

8. Pharoah PD, Dunning AM, Ponder BA, Easton DF. Association studies for finding cancer-susceptibility genetic variants. Nat Rev Cancer 2004;4:850–60.
9. Ionnidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ionnidis DG. Replication validity of genetic association studies. Nat Genet 2001;29:306–9.
10. Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. Lancet 2003;361:865–72.
11. Brennan P. Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? Carcinogenesis 2002;23:381–7.
12. Rebbeck TR, Martinez ME, Sellers TA, Shields PG, Wild CP, Potter JD. Genetic variation and cancer: Improving the environment for publication of association studies. Cancer Epidemiol Biomarkers Prev 2004;13:1985–6.
13. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004;96:434–42.
14. Thomas DC, Clayton DG. Betting odds and genetic associations. J Natl Cancer Inst 2004;96:421–3.
15. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A 1999;96:2907–12.
16. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 2001;69:138–47.
17. Sillanpaa MJ, Corander J. Model choice in gene mapping: what and why. Trends Genet 2002;18:301–7.
18. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 2003;4:701–9.
19. Tahri-Daizadeh N, Tregouet DA, Nicaud V, Manuel N, Cambien F, Tiret L. Automated detection of informative combined effects in genetic association studies of complex traits. Genome Res 2003;13:1952–60.
20. Cook NR, Zee RY, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. Stat Med 2004;23:1439–53.
21. Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. Bioinformatics 2004;20:1896–904.
22. Millstein J, Siegmund KD, Conti DV, Gauderman WJ. Identifying susceptibility genes by using joint tests of association and linkage and accounting for epistasis. BMC Genet. In press 2005.
23. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. Stat Med 1993;12:717–36.
24. Witte JS. Genetic analysis with hierarchical models. Genet Epidemiol 1997;14:1137–42.
25. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Stat Sci 1999;14:382–417.
26. Viallefont V, Raftery AE, Richardson S. Variable selection and Bayesian model averaging in case-control studies. Stat Med 2001;20:3215–30.
27. Cortessis V, Thomas DC. Toxicokinetic genetics: an approach to gene-environment and gene-gene interactions in complex metabolic pathways. In: Bird P, Boffetta P, Buffler P, Rice J, editors. Mechanistic considerations in the molecular epidemiology of cancer 157. Lyon (France): IARC Scientific Publications; 2003. pp. 127–50.
28. Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. Hum Hered 2003;56:83–93.
29. Clewell HJ, Andersen ME, Barton HA. A consistent approach for the application of pharmacokinetic modeling in cancer and noncancer risk assessment. Environ Health Perspect 2002;110:85–93.
30. Gelman A, Bois F, Jiang J. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. J Am Stat Assoc 1996;91:1400–12.
31. Wakefield J. The Bayesian analysis of population pharmacokinetic models. J Am Stat Assoc 1996;91:62–75.
32. Tonolio P, Boffetta P, Shuker DEK, Rothman N, Hulka B, Pearce N. Application of biomarkers in cancer epidemiology. Lyon: IARC Scientific Publications; 1997.
33. Potter JD. Toward the last cohort. Cancer Epidemiol Biomarkers Prev 2004;13:895–7.
34. Gunter MJ, Probst-Hensch NM, Cortessis VK, Kulldorff M, Haile RW, Sinha R. Meat intake, cooking-related mutagens and risk of colorectal adenoma in a sigmoidoscopy-based case-control study. Carcinogenesis 2004.
35. Stram D, Longnecker M, Shames L, et al. Cost-efficient design of a diet validation study. Am J Epidemiol 1995;142:353–62.
36. Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. Biometrics 1991;47:851–69.
37. Hung RJ, Brennan P, Malaveille C, et al. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. Cancer Epidemiol Biomarkers Prev 2004;13:1013–21.
38. Gilliland FD, Li YF, Saxon A, Diaz-Sanchez D. Effect of glutathione-*S*-transferase M1 and P1 genotypes on xenobiotic enhancement of allergic responses: randomised, placebo-controlled crossover study. Lancet 2004;363:119–25.
39. Waters MD, Fostel JM. Toxicogenomics and systems toxicology: aims and prospects. Nat Rev Genet 2004;5:936–48.
40. De Roos AJ, Smith M, Channock S, Rothman N. Toxicologic considerations in the application and interpretation of susceptibility biomarkers in epidemiologic studies. In: Bird P, Boffetta P, Buffler P, Rice J, editors. Mechanistic Considerations in the Molecular Epidemiology of Cancer (this volume). Lyon: IARC Scientific Publications. No. 157; 2004. pp. 105–25.
41. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002;30:3894–900.
42. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4.
43. Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974;185:862–4.
44. Xi T, Jones IM, Mohrenweiser HW. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics 2004;83:970–9.
45. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 2003;33:228–37.
46. Savas S, Kim DY, Ahmad MF, Shariff M, Ozcelik H. Identifying functional genetic variants in DNA repair pathway using protein conservation analysis. Cancer Epidemiol Biomarkers Prev 2004;13:801–7.
47. Zhu Y, Spitz MR, Amos CI, Lin J, Schabath MB, Wu X. An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. Cancer Res 2004;64:2251–7.
48. Game JC, Birrell GW, Brown JA, et al. Use of a genome-wide approach to identify new genes that control resistance of *Saccharomyces cerevisiae* to ionizing radiation. Radiat Res 2003;160:14–24.
49. Kass R, Raftery A. Bayes factors. J Am Statist Assoc 1995;90:773–95.
50. Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO. Metabolic pathways in the post-genome era. Trends Biochem Sci 2003;28:250–8.
51. Westerhoff HV, Palsson BO. The evolution of molecular biology into systems biology. Environ Health Perspect 2004;22:1249–52.
52. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 2000;7:601–20.