Cancer Risk Assessment at the Atomic Level

Alvaro N.A. Monteiro¹ and Fergus J. Couch²

¹Risk Assessment, Detection, and Intervention Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida and ²Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota

Abstract

A flurry of articles on the structure of different domains of BRCA1 and BRCA2 have not only shed light on the biology of these proteins but have also raised hopes that these data could eventually be used to infer cancer association for a large number of inherited missense mutations whose effect on protein function is unclear. (Cancer Res 2006; 66(4): 1897-9)

A key to an effective cancer prevention strategy is the early identification of individuals at elevated risk of disease. Thus, individuals found to carry inactivating mutations in tumor suppressor genes, exemplified in the case of the two major breast cancer susceptibility genes, BRCA1 and BRCA2, can benefit from presymptomatic interventions, such as surgery and chemoprevention (1). However, whereas a majority of people undergoing testing will receive informative results, a significant portion (~13%) are left with an often enigmatic diagnostic report: genetic variant of uncertain significance, also termed unclassified variant (2). Unclassified variants include mainly missense mutations but also include mutations in regulatory regions and synonymous mutations in splicing enhancer regions for which the effect on protein function has not been determined. Unclassified variants now account for 40% of all sequence alterations excluding common polymorphisms that are identified by mutation screening of BRCA1 and BRCA2. Importantl, there is also emerging evidence that the percentage of individuals receiving noninformative results is significantly higher in minority populations. In a prescient comment during the early days of BRCA research, Francis Collins suggested that from the perspective of screening, these missense mutations were going to plague us (3). Indeed, unclassified variants, which can also be found in other cancer predisposing genes, such as MLH1, MSH2, and ATM, represent a major clinical issue, in that the inability to tell an individual carrying an unclassified variant whether the mutation is cancer predisposing or not constitutes a significant problem for risk assessment, genetic counseling, and informed decision making about cancer prevention and therapeutics.

A straightforward conclusion is that a thorough understanding of protein function is badly needed before the effect of unclassified variants on protein activity can be assessed and improved counseling can be provided to patients. Recently, a series of approaches based on sequence analysis have been used to predict the possible effect of unclassified variants on protein function (4–6). However, these methods are limited by their inability to account for the three-dimensional architecture of proteins. In the last decade or so, determination of the three-dimensional structure of proteins has been instrumental in revealing important aspects of protein function, such as the regulation of protein kinases and ion channels (7, 8). Germane to the case in point, the availability of solution and crystal three-dimensional structures of several different domains of BRCA1, BRCA2, and other related proteins (9–21) has the potential to be a watershed event in cancer risk assessment. Through a detailed understanding of structure-function relationships, we should be able to generate reliable computation prediction methods for assessment of risk at the atomic level for unclassified variants. Although one should be hopeful that cancer risk assessment at the atomic level will be a reality, there are significant hurdles to be overcome that will take a close collaboration of structural and molecular biologists, epidemiologists, computer scientists, geneticists, and genetic counselors. What then are the hurdles that need to be overcome to achieve this goal?

Misclassifying the Unclassified: Overcoming the Two Cultures Problem

In a classic example of C.P. Snow’s Two Cultures problem, interdisciplinary approaches create the imperative need not only to adapt the language but also to convey, across discipline borders, the exact weight of each type of evidence. When perusing the BRCA1 and BRCA2 literature, we encounter the common usage of terms such as “tumor-derived,” “cancer-associated,” “clinically-relevant,” and “cancer-linked” to refer to missense changes in many cases implying that disease association has been firmly established. In fact, whereas a number of variants has been unequivocally associated with cancer, most of the missense variants deposited in the Breast Cancer Information (BIC) BRCA1 and BRCA2 mutation database lacks enough information for determination of their association with cancer predisposition. Mutations unequivocally associated with cancer predisposition have been characterized predominantly through assessment of available genetic and family data. Mutations deposited in the database that are not explicitly classified as deleterious represent variants that have been found in a specific biological sample (which is rarely tumor-derived as most of them are detected as germ line mutations present in samples from blood). They have not (as yet) been formally associated with disease using genetic approaches. Thus, the correlation between structure and function has been over interpreted through misunderstandings about the nature of the unclassified variants in the database. Another common misconception is equating frequency in the database with allelic frequency in the population. Submission to the database is entirely voluntary, and depositors have a tendency to not report a mutation that has already been deposited (as it is not novel!), not realizing that this...
decision increases the disconnect between the frequency in the database and the allelic frequency in various populations. It is important to note that BRCA1 and BRCA2 mutations and by extension entries in the database vary widely by origin and population-specific frequency. Thus, use of frequency data from the BIC database to identify an unclassified variant as disease predisposing is inappropriate. To avoid these misconceptions, curators and administrators of mutation databases must realize that their audience is progressively larger and more diverse. To better serve the research, clinical, and consumer communities the descriptive terms, categories and criteria for classification of the disease status must be clearly outlined and should not rely on in-depth knowledge of the genetics of the disease on the part of the user.

Separating the Wheat from the Chaff: Identifying Informative Mutations

Another pervasive problem is the use of certain mutations to decide whether a particular protein function is important for cancer. In many cases, investigators studying a particular biochemical function want to determine if the function in question is related to the tumor-suppressive action of the protein. For that, they will identify a mutation that has been found, by genetic methods to be unequivocally associated to cancer, introduce it in the protein and assay for the abrogation of function. If the function is abrogated, then it must be important for cancer. Unfortunately, this is not necessarily so. For example, one can imagine that a particular domain mediates two different functions, only one of which is important for the tumor-suppressive action. However, introduction of a mutation that will completely disrupt folding will necessarily abolish both functions. The investigator in question, having assayed a favorite function will conclude, incorrectly, that the function is required for the tumor-suppressive actions of the protein. As structural models can be used to predict such severe consequences, these models should prove invaluable in predicting which unclassified variants substantially alter structure and function and should serve to identify a subset of unclassified variants to be further evaluated by genetic models or other methods. Furthermore, the availability of the structures and models now allow us to identify mutations that do not disrupt folding but rather affect specific binding sites.

Reproducibility and Validity

To create reliable computation prediction methods, the results from functional analyses that are going to be correlated with structural variables must be derived from valid and reproducible assays. This is true whether the assay is the measure of a bona fide function or a surrogate test that is able to monitor the structural integrity of the protein. Current assays have been conducted in a research setting with no systematic approach designed to test specificity (the probability that a test will be negative when administered to alleles not correlated with disease) and sensitivity (the probability that a test will be positive when administered to alleles correlated with disease) or to a rigorous examination of internal controls. Importantly, structure-function analyses are only as good as their controls. Investigators involved in the development of these assays need to be upfront about the assays’ limitations as they are hardly comprehensive and many proteins, such as BRCA1 and BRCA2, will have multiple biochemical functions potentially contributing to the phenotype.

What If There Is No Available Structural Information?

Although there has been enormous progress in producing high-quality three-dimensional structures (22), there is still a large number of proteins and protein regions for which there is no structure information. This is illustrated by portions of BRCA1 and BRCA2, for which no structural information exists. Although part of this problem may eventually be resolved using de novo prediction methods (in which structural prediction is not based on a known structural template), one needs to consider regions that are intrinsically unstructured as has been indicated recently for BRCA1 (23, 24). In this case, several comparative evolutionary methods that do not rely on structure but rather on primary sequence conservation in a series of orthologues have provided important information for the evaluation of BRCA1 variants and hold tremendous potential for the analysis of unstructured regions (4, 5, 25).

Penetrance and Protein Activity: a Two-Layer Problem

One of the important unknowns today in classifying tumor suppressor gene alleles is related to the nature of how protein activity relates to penetrance. On one side, penetrance may be correlated to protein activity in a continuous manner: the lower the activity, the higher the risk over a wide range of activities. If this is the case, quantitative tests for protein function must be devised. Alternatively, the correlation may be discrete. In this case, no matter how low the activity is, unless it crosses a minimum activity threshold risk will be the same. Clearly, this is an important problem because in the former scenario alleles need to be classified according to a quantitative measurement of risk, whereas in the latter, alleles can be qualitatively classified into either high or low risk (low being the risk in the general population). In addition, this reveals the need to identify the minimum activity allele required to confer tumor-suppressive activity in the organism. Although many biological systems show robustness, it is not clear if the different predisposition alleles will present similar penetrances.

The relationship between protein activity and penetrance is only part of the problem as penetrance may also be highly influenced by genetic background. Determination of the influence of other loci impinging on the activity of a tumor suppressor gene is still a fleeting goal at this point, but genome-wide single nucleotide polymorphism (SNP) analysis might eventually yield some clues. This approach combined with the use of generalized computation prediction methods to predict the effect of missense changes on a large set of proteins is a promising avenue that carries its own potential pitfalls discussed below.

Generalization: Avoiding the Catch-22

Molecular biology approaches are eminently reductionist, and critics have been relentless in pointing out that these approaches have their days numbered. This valid critique assumes, perhaps wrongly, that generalizations will be necessarily better. General computational prediction methods that are based on the universe of known proteins and mutants have the tremendous advantage of a wealth of data that can be used to allow fine tuning of threshold values used in specific features (26–28). These approaches rely on mapping a subset of common variants, the nonsynonymous SNPs, to structural surfaces to predict the effect of the amino acid replacement on the function of the protein. Although these
approaches have been successful in predicting severe effects and in deriving general rules (e.g., a correlation of deleterious variants and sites of low solvent accessibility; ref. 27), it is not clear how effective they will be in identifying subtle changes in binding surfaces. Tumor suppressor genes often encode proteins with complex and multiple functions in which protein binding surfaces are multiple and often still unknown. Surface mutations with little effect on folding but that disrupt a critical binding interaction may not be correctly evaluated by general computational prediction methods. On the other hand, individualized computational prediction methods based on mapping rare allelic variants of a single locus to one structural surface may incorporate features that take into account binding surfaces and other protein-specific characteristics. Unfortunately, these approaches are generally heuristic and suffer from a paucity of data points. This dearth of data precludes regression approaches leading to a high number of variants, for which there is not enough information to decide whether it will affect function or not (29). Genome-wide analysis of the effect of coding SNPs using individualized prediction methods is presently impractical, but there is no technical impediment to this becoming a reality. The challenge will be to design generalized core computational methods that have add-ons, taking into account protein interactions and specialized functions. Although this may not lead to an elegant approach, it is practical.

Changing the Face of Clinical Practice

It is a logical conclusion that the fundamentals that will be established during this process of detailed structure-function analysis will be a treasure chest for several areas of clinical practice. How would this future look like? In the past, classic methods of epidemiology were extremely successful in identifying groups of individuals at high risk for cancer. The cloning of tumor suppressor genes, such as TP53 and APC, in familial syndromes, such as Li-Fraumeni and familial adenomatous polyposis, respectively, ignited the hope of making the transition from the determination of risk in a group of individuals to individual cancer risk (30). This hope was dampened by the realization that there was a wide spectrum of mutations that varied in their penetrance and specific cancer phenotype. The structure-based approaches could in principle provide this much needed beaconhead in a frontal attack to determine individual risk in that the individual genetic variation could be correlated with one’s protein activity and subsequently to individual risk of cancer. A likely scenario will be that this information on protein activity and phenotype may also allow a precise determination of drug response or treatment outcome. In addition, recent large sequencing efforts to identify somatic mutations in tumors have generated a throve of sequence variants (31, 32). The expansion of these efforts will generate a large number of somatic mutations, for which the effect on function will be difficult to determine. The structural-based approach will also be important to face this challenge. We now have the basis to start a systematic evaluation of how every mutation may affect protein function and how it will eventually influence the organism’s phenotype.

Acknowledgments

Received 8/25/2005; revised 11/28/2005; accepted 12/14/2005.

Grant support: Work in the Monteiro and Couch Labs is supported by the NIH grants CA92309 (A. Monteiro) and CA82267, an American Cancer Society Research Scholar Award, and the Breast Cancer Research Foundation (J.J. Couch). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Tom Sellers, Rachel Karchin, Cathy Phelan, and members of the Monteiro Lab for helpful comments.

References


www.aacrjournals.org

Cancer Risk Assessment at the Atomic Level

Received 8/25/2005; revised 11/28/2005; accepted 12/14/2005.

Grant support: Work in the Monteiro and Couch Labs is supported by the NIH grants CA92309 (A. Monteiro) and CA82267, an American Cancer Society Research Scholar Award, and the Breast Cancer Research Foundation (J.J. Couch).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Tom Sellers, Rachel Karchin, Cathy Phelan, and members of the Monteiro Lab for helpful comments.
Cancer Risk Assessment at the Atomic Level

Alvaro N.A. Monteiro and Fergus J. Couch