Integration of structural dynamics and molecular evolution via protein interaction networks: a new era in genomic medicine

Avishek Kumar¹,⁵, Brandon M Butler¹,⁵, Sudhir Kumar²,³,⁴ and S Banu Ozkan¹

Sequencing technologies are revealing many new non-synonymous single nucleotide variants (nsSNVs) in each personal exome. To assess their functional impacts, comparative genomics is frequently employed to predict if they are benign or not. However, evolutionary analysis alone is insufficient, because it misdiagnoses many disease-associated nsSNVs, such as those at positions involved in protein interfaces, and because evolutionary predictions do not provide mechanistic insights into functional change or loss. Structural analyses can aid in overcoming both of these problems by incorporating conformational dynamics and allostery in nsSNV diagnosis. Finally, protein–protein interaction networks using systems-level methodologies shed light onto disease etiology and pathogenesis. Bridging these network approaches with structurally resolved protein interactions and dynamics will advance genomic medicine.

Addresses
¹ Department of Physics and Center for Biological Physics, Arizona State University, Tempe, AZ 85281, United States
² Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122, United States
³ Department of Biology, Temple University, Philadelphia, PA 19122, United States
⁴ Center for Genomic Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia

Corresponding author: Ozkan, S Banu (Banu.Ozkan@asu.edu)
⁵ These authors contributed equally to this work.

Introduction

Proteins are the remarkable workhorses of life, as they play crucial roles in biological function. They carry out their function through complex, carefully orchestrated protein–protein interactions in a crowded cellular environment. There have been many efforts to understand living systems by identifying protein interactions, including high-throughput methods such as yeast two-hybrid systems [1–3] and high affinity purification followed by mass spectrometry [4]. Moreover, these experimental efforts have been combined with computational approaches, making it possible to generate protein–protein interaction (PPI) networks at different genomic scales, from metabolic pathways to a diversity of species from bacteria to humans [5].

In addition to the tremendous amount of data arising from PPI networks, another front has emerged through genomic sequencing. For the past two decades, scientists have been profiling genomic variations in healthy and diseased individuals. Genome-wide association studies, whole-genome sequencing, and exome sequencing have shown that each personal genome contains millions of variants, thousands of which are non-synonymous single nucleotide variants (nsSNVs). Many of these nsSNVs are associated with Mendelian and complex diseases [6]. With the sequencing of each new personal exome, the constellation of nsSNVs is expanding at a fast rate. But the translation of a personal exome variation profile into biomedically relevant information remains a challenge, particularly because a large proportion of novel nsSNVs are rare [7].

In this review, we discuss approaches for diagnosing the potential disease functional impact of nsSNVs (Figure 1). First, we review methods based on detailed evolutionary and biophysical information, where molecular structures of protein complexes and the corresponding conformational dynamics information are utilized. Then, we review systems-level approaches of PPI networks to identify disease-associated mutations and disease pathology. A unified approach that merges these three major levels of information in diagnosing benign and disease-associated nsSNVs can provide solutions to the current challenges in genomic medicine [8*,9*].

Evolutionary and structural approaches for prediction of disease mutations

A large number of computational tools employ purely evolutionary information to predict the impact of nsSNVs, under the auspices of the neutral theory of molecular evolution [10*,11]. Simply put, evolutionarily permissible substitutions in the amino acid sequence are determined by comparing sequence homologs across the
Computational tools in genomic medicine — two different approaches on different scales are used to determine the impact of nsSNVs on protein–protein interactions. (a) Structural methods use the information provided by hotspots [40,51] and estimate the change in binding affinity upon mutation. Hotspots (red spheres) are functional residues that are crucial for binding in protein–protein interactions, while non-hotspot residues (wheat spheres) still participate in interactions at the interface but are not crucial. If a hotspot is mutated, then it will result in a deleterious mutation. Evolutionary metrics leverage positional conservation, which is determined by a multiple sequence alignment of many different species across the tree of life to measure the effect of nsSNVs [11]. Mutations at conserved sites are usually deleterious. The phylogenetic tree shown here presents the 10 mutations associated with Miller syndrome, a rare genetic disorder. All mutations occur at slowly evolving or highly conserved sites both in primates and distantly related vertebrates. (b) Protein–protein interaction (PPI) networks represent a systems-level approach where each node represents a protein and undirected edges represent their interactions. Disease mutations leading to a loss of an interaction can be shown as a removal of an edge, while knock out mutations can be captured as a node removal. Local and global network metrics can be used to find disease-associated variants.

The evolution of diverse species. If an nsSNV is not found in the observed variation across the phylogeny, then it may be diagnosed to be putatively disease-associated (i.e., function impacting). To be more precise, probabilistic scoring functions are developed by using amino acid positional conservation and molecular phylogenetics. Current evolution-based diagnosis methods are widely used and are considered to produce good estimates [11–21]. However, they do have blind spots [11] and their accuracies in practical applications are debated because of their need to use training data that may not reflect the distribution of nsSNVs in the application domain [22–24].

Some of the current approaches combine evolutionary considerations with structural information in order to improve the prediction accuracy [21,25–27]. For instance, PolyPhen-2 uses solvent accessibility, secondary structure propensity, and crystallographic B-factors to classify mutational sites [21]. Other approaches consider the change in polarity, volume, and charge of the amino acid. Solvent accessibility has been used in a number of phenotypic prediction studies and has proven to be a useful attribute in disease prediction [26]. Moreover, residue–residue interaction networks of protein structures are used to identify functionally important residues through network topology parameters [27,28], and are utilized in predicting the impacts of observed nsSNVs. While the evolution-based methods are more effective than methods that solely use structural features, their accuracy breaks down at less-conserved positions resulting in true positive rates less than 50% [11,29]. These methods also have great difficulty in diagnosing benign variations at
highly conserved positions (<50% rate of correct diagnosis of true negatives) [30]. It has also been shown that in silico tools yield very low accuracy for nsSNVs found to be associated with complex diseases; PolyPhen-2 [29] produced 22% true positives for 757 variants from VARMED [31]. This low accuracy is due to multiple genes having small cooperative effects in complex diseases whose disease-associated nsSNVs are often not located at highly conserved positions [32,33,34,35].

Beyond evolutionary conservation, there have been many efforts to utilize structural and network properties to diagnose disease variants. An all-atom structural mapping of observed nsSNVs on human PPI networks revealed that disease-associated nsSNVs are largely found at protein–protein interfaces [34**]. For this reason, some recent methods have focused on modeling interfaces and predicting changes in binding affinities to distinguish the disease-associated nsSNVs from neutral nsSNVs. The proliferation of available experimental structures in the Protein Data Bank [36] and current advancements in homology modeling have facilitated the development of human structural interaction network (HSIN) databases of protein–protein and domain–domain interactions [37]. Mapping neutral and disease-associated nsSNVs on HSIN has shown several important results [32,33,34**,38,39]. First, these studies showed that the pleiotropy of disease-associated nsSNVs can be explained by proteins interacting with different proteins at different interfaces [33], where the mutations at these separate interfaces may lead to different diseases and intensities [34**]. Second, nsSNVs at interfaces may disrupt or enhance protein–protein interactions, thus, playing an important role in pathogenesis [38,40]. While the disruption of transient binding interactions can usually affect the protein localization, the loss of stable interactions due to interface mutations leads to complete loss of function. The mutations that enhance binding interactions may cause aggregation or aberrant recognition, as observed in cancers [39].

Because interface mutations may alter binding interactions, there have been efforts to predict the effects of these mutations by measuring the difference between the free-energy change upon binding of the wild type and mutant (ΔΔG). Free energy differences upon binding calculated via thermodynamic integration and free energy perturbation approaches combined with molecular dynamics simulations are computationally expensive, particularly for large-scale protein complexes [41]. Therefore, many have developed in silico tools as a fast alternative to estimate ΔΔG using statistical energy functions based on known protein structures [42–44] and/or coupling with machine learning tools using training sets [45–47]. However, these calculations can be rather inaccurate, because local structural changes upon mutations are generally neglected [48,49].

Teng et al. used an all-atom molecular force field (CHARMM) to investigate the effect of disease and neutral nsSNVs on binding energies for 264 protein–protein complexes with known nsSNVs. They found that disease-associated mutations often destabilize the electrostatic component of the binding energies. Furthermore, the change in physicochemical properties upon mutation, such as large changes in polarity and hydrophobicity, do not significantly alter the binding energy, making it challenging to distinguish between disease and benign nsSNVs [50]. Evaluating the importance of a particular interface residue to binding is another approach to predict the impact of nsSNVs.

Experimentally, crucial binding sites can be identified by mutating each site to alanine and measuring the change in binding affinity. These positions, called hotspots, are often located at highly conserved positions with large changes in accessible surface area (ASA) upon binding [8*,51]. If a mutation occurs on such a site, it will impact function and, possibly, be deleterious. Incorporating biophysical and structural properties of known hotspots into machine learning algorithms have made it possible to distinguish between disease-associated and neutral nsSNVs at protein interfaces [38]. It remains a challenge to predict disease-associated mutations occurring at non-hotspots.

Conformational dynamics and allosterity in disease development

Currently, most machine learning methods that use structural features (e.g., ASA) are based on static 3D structures. This practice neglects protein conformational dynamics. However, protein structure-encoded conformational dynamics, which span a broad timescale of motion from atomic fluctuations and side chain rotations to collective domain movements, underlie a protein’s biological function. Protein evolution studies of several different protein families have shown that changes in conformational dynamics through allosteric regulation lead to new functions (e.g., green fluorescent protein (GFP), beta-lactamase inhibitors, and nuclear receptors [52–54]). Moreover evolutionary rates are strongly correlated with the flexibility of individual positions obtained from conformational dynamics [55–57].

Protein dynamics studies assert that protein function can be explained by analyzing the individual contribution of residues to the conformational dynamics and stability of a protein [55,56,58*]. Therefore, conformational dynamics-based metrics can also be utilized in predicting the impact of nsSNVs on protein function. Gerek et al. used an amino acid site-specific dynamic flexibility index (DFI) metric to evaluate the effect of flexibility of individual positions on biological fitness and function. DFI is a position-specific metric that quantifies the resilience of each residue to a perturbation occurring at another part of the chain, thus identifying the flexible and rigid parts
of a protein [55]. Analysis of disease-associated and neutral nsSNVs for more than 100 human proteins revealed that disease-associated nsSNVs occur predominately at low DFI sites (i.e., rigid hinge sites), signifying the importance of hinge sites that control functionally crucial motions. In contrast, neutral variants are more abundant at positions with high DFI, suggesting that flexible sites are more robust to mutations [55]. Furthermore, DFI profiles of over a thousand positions harboring mutations revealed that positions at protein interfaces have lower average DFI than those at non-interfaces, suggesting that protein–protein interfaces have less dynamic flexibility [58]. These results suggest that hinge positions at interfaces are crucial for binding and mutations at these hinge sites will probably lead to disease.

Allostery is the regulation of cellular functions through the alteration of protein dynamics and structure upon an action at a distant site, which has been implicated in diseases. There are several disruptions of allosteric regulations that lead to disease development. Mutations can allosterically impair post-translational modification as observed in driver mutations in cancer [59,60,61]. Disease-associated variances can also change the ON/OFF populations in cell signaling by altering the stability of certain conformations and/or dynamics. Furthermore, they can lead to disease by shifting allosteric pathways, as observed in the mutation that gives rise to hyperekplexia [62]. Finally, mutations farther away from functionally crucial sites can allosterically impair hinges (i.e., rigid parts), softening the functionally crucial regions and lead to the loss of allosterically regulated conformational dynamics as observed for disease-associated mutations of human ferritin [63†].

Allostery can elucidate the impact of non-hotspot mutations dynamically linked to hotspots [7]. Hotspots evaluated by the HotPoint server [51] of the protein assemblies in the dataset studied by Butler et al. [58†] indicated that most mutations occurring at hotspots are disease-associated. However, among the hundred disease-associated nsSNVs at interfaces, only half of them were at hotspots. How do non-hotspot sites play a role in disease-assocation? This can be studied by a new metric called the functional dynamic flexibility index (f-DFI) [63†]. f-DFI quantifies the resilience of each residue upon the perturbation of a functionally crucial distant site. Thus, f-DFI enables the identification of non-hotspots residues that are linked allosterically to hotspots. Interestingly, ~80% of disease-associated mutations at non-hotspots exhibited high f-DFI values (>0.6), indicating they are dynamically coupled to hotspot residues. Figure 2 presents a case study of two protein complexes, alanine:glyoxylate aminotransferase and lysosomal beta-hexosaminidase A. In this case, benign mutations have low f-DFI (<0.4), despite being in close proximity to hotspots. In contrast, the disease-associated mutations have high f-DFI (>0.6), indicating they are dynamically linked to hotspots. Two are spatially close to a hotspot, so it is not surprising that they have high f-DFI scores. However, the others are

Figure 2

Case study using f-DFI for mutations on two protein complexes — (a) two protein complexes used in a previous study [58†], alanine:glyoxylate aminotransferase (AGT) and (b) lysosomal beta-hexosaminidase A (Hex A), PDB codes 1h0c and 2gk1, respectively. Both are enzymes in the human body that are responsible for functions in the liver and central nervous system. One chain on each dimer is color-coded by its f-DFI score (0 to 1) for each residue, where blue indicates low f-DFI and red high f-DFI [63†]. On the left AGT displays 3 neutral mutations, while on the right Hex A shows 4 known disease-associated mutations purported by the Human Gene Mutation Database (HGMD) [87]. None of the mutations are located at hotspots, making them difficult to predict. f-DFI is a metric that can capture the dynamic coupling between mutations at non-hotspots and hotspots despite not being in close proximity.
not as close, they are dynamically coupled with hotspots making them crucial sites. When non-hotspot sites dynamically coupled to hotspots are mutated, loss of function may occur and result in a potentially detrimental phenotype.

**Network metrics can identify disease-associated proteins in PPI networks**

Beyond the joint evolutionary and structural analysis of single proteins, the diagnosis of disease causing nsSNVs for complex diseases would require the analysis of multiple proteins together that are connected in PPI networks. In PPIs, proteins and their interactions are represented as nodes connected by undirected edges [64–66, 67] without taking into account the details of molecular interactions. PPI networks are described by scale-free networks having hubs with a high degree of connectivity; thus, they have the important property of being resilient to random stochastic effects, a necessary property in biology [68]. Disease can manifest itself in two ways in networks: node removal or edge modification. When a node is removed from a network, it is due to a destabilizing mutation that knocks out a protein. An edge modification is due to removing or adding an edge in the interaction network. It has been experimentally shown that many edgetic mutations are due to mutations on interfaces [69]. Edges (interactions) can also be added, leading to gain-of-function mutations [70].

Local and global network metrics combined with known disease-associated proteins can reliably predict unknown disease-associated proteins. The first attempts to identify disease-associated nsSNVs in PPI networks used local metrics such as the Direct Neighbor Counting method (also known as the guilt-by-association method), where it is assumed that candidate proteins that interact with known disease proteins are themselves disease-associated [71, 72]. Global metrics can identify disease proteins that do not directly interact with known disease proteins [67]. In the shortest path analysis method, the shortest path between two disease nodes is found. A node in close proximity to multiple disease nodes has a high probability of being disease-associated [73]. It has also been shown that ‘bottleneck’ or ‘sole-broker’ proteins with a high betweenness/centrality (i.e., many shortest paths passing through a node) are also probably to be disease-associated [74, 75]. Methods such as diffusion kernel and random walk with restart measure how two non-interacting nodes are related by having random walkers start from a known disease node and diffuse through the network [76]. These global metrics enable the identification of the nodes and edges that are associated to known disease genes by exploiting the full network topology.

Proteins that interact with several disease proteins or proteins that are in proximity to disease proteins will have a higher probability to be encoded by a disease gene [76]. Köhler et al. showed that random walk with restart is superior to local metrics [76]. Although random walk methods produce the most accurate results, they still fail to identify disease variants predicted by local methods. Navlakha and Kingsford were able to create a consensus method using 13 different metrics in tandem in an ensemble of decision trees with a random forest classifier [77]. This method currently has the best accuracy. By incorporating multiple-omics (e.g., genomics, transcriptomics, and proteomics) analysis into network methods, Chen et al. were able to identify biological processes for two viral infections and the development of type 2 diabetes [78]. The robustness of the PPI network used is crucial for the higher accuracy of these approaches [79–82]. Guney and Oliva [83] tested several network-based methods with respect to the perturbations of the system using various disease phenotypes from the Online Mendelian Inheritance in Man (OMIM) database. They found that disease proteins are connected via multiples pathways in a PPI network. Even when these networks are significantly perturbed, network-based methods can reveal hidden disease association proteins, particularly in cases of breast cancer and diabetes. In general, the PPI network approaches can identify certain proteins associated with specific disease better than others [77, 83].

Overall PPI networks represent the simplest networks. They capture whether proteins interact, the architecture of a network, but they do not tell us how or at what rate they interact, and what the parameters of the network are. There have been cases studies where interactions in PPI networks can be parameterized by rate constants [84–86]. Because of difficulties of measuring parameters in a cellular context, parameterization of a proteome wide PPI network in humans has yet to be realized. Thus, the future of network approaches in PPI analysis lies in creating more accurate PPI datasets and integration of different omics.

**Conclusion**

Advances in sequencing technologies are providing a myriad of data on human genetic variation. However, distinguishing between neutral variants (with little or no effect on phenotype) from variants conferring disease risk remains elusive. While earlier methods did not consider the role of protein interactions in the identification of disease-associated variants, recent studies about the prevalence of nsSNVs at interfaces provided mechanistic insight about their crucial role in interactions. This has led to two different approaches at different length scales: PPI networks at the system level and biophysical methods and evolutionary information at the molecular level. The future in genomic medicine lies in merging these two approaches. The information of how two proteins interact in a PPI network, rather than merely knowing two proteins interact, will provide the next major advance-
ment to uncover disease pathology of Mendelian, particularly, complex diseases. As PPI data improves and new nsSNVs are discovered, we will enter a new phase of genomic medicine.

Conflict of interest statement
Nothing declared.

Acknowledgements
Support from NIH awards U54HG0084599 and LM011941-01 is gratefully acknowledged by SBO. SK also acknowledges HG002096-12, LM011941-02, and DK098242-03. We thank Dr. Ashini Bolia for a careful review of the manuscript.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:
• of special interest
• • of outstanding interest
A comprehensive review on protein–protein interaction networks.
The authors detail the use of structural networks that move beyond the PPI network paradigm. In structural networks the types of interactions between proteins are described as well as the nature of how proteins bind with one another. This is in contrast to PPI edge and node networks that only describe which protein interact with others.
A perspective on the null hypothesis for evaluating disease-associated variation, which ultimately enables one to use the principles of neutral theory of molecular evolution in diagnosing disease nsSNVs.
30. Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ: Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and


61. The authors show how dynamics can distinguish between neutral and disease associated mutations in human ferritin protein by allosterically changing the crucial functional hinge sites.


66. A comprehensive review of network approaches using networks to predict and understand the mechanics of human disease.


This work shows the power of integrating multiple omics approaches (genomics, transcriptomic, proteomic, and metabolomic) to analyze medical risks, such as type 2 diabetes, and dynamics changes in biological pathways between healthy and diseased individuals.


