Advances in Human Biology: Combining Genetics and Molecular Biophysics to Pave the Way for Personalized Diagnostics and Medicine

Emil Alexov

Computational Biophysics and Bioinformatics, Department of Physics, Clemson University, Clemson, SC 29634, USA

Correspondence should be addressed to Emil Alexov; ealexov@clemson.edu

Received 29 January 2014; Revised 23 April 2014; Accepted 17 June 2014; Published 7 July 2014

Academic Editor: Yinan Wei

Copyright © 2014 Emil Alexov. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Advances in several biology-oriented initiatives such as genome sequencing and structural genomics, along with the progress made through traditional biological and biochemical research, have opened up a unique opportunity to better understand the molecular effects of human diseases. Human DNA can vary significantly from person to person and determines an individual's physical characteristics and their susceptibility to diseases. Armed with an individual's DNA sequence, researchers and physicians can check for defects known to be associated with certain diseases by utilizing various databases. However, for unclassified DNA mutations or in order to reveal molecular mechanism behind the effects, the mutations have to be mapped onto the corresponding networks and macromolecular structures and then analyzed to reveal their effect on the wild type properties of biological processes involved. Predicting the effect of DNA mutations on individual's health is typically referred to as personalized or companion diagnostics. Furthermore, once the molecular mechanism of the mutations is revealed, the patient should be given drugs which are the most appropriate for the individual genome, referred to as pharmacogenomics. Altogether, the shift in focus in medicine towards more genomic-oriented practices is the foundation of personalized medicine. The progress made in these rapidly developing fields is outlined.

1. Introduction

The human body is a delicate, self-regulating machine which can respond to its surroundings and internal needs. Such self-regulation involves various processes ranging from processes on atomic and molecular level to processes occurring in organs and tissues. Despite such tremendous complexity, somehow all humans, broadly speaking, are quite similar. However, slight differences in DNA can lead to a multitude of other physical differences. Some of these differences are harmless such as eye and hair color [1], race [2], and skin color [3, 4], while other differences may be disease-associated (see special J. Mol. Biol. issue [5]). The differences among individuals and their susceptibility to diseases are not only due to the single nucleoside polymorphisms (SNPs), but also due to the fact that different individuals have different copy numbers variations (CNVs) for various genes [6–9]. As pointed out by Haraksingh and Snyder [6], the CNVs are perhaps even more important for the humans than the SNPs, a statement supported by other researchers [10–13]. In the end, from the viewpoint of personalized diagnostics and medicine, the most important task is to differentiate between disease-causing and harmless DNA differences. At the same time, from the viewpoint of Biology and Biophysics, one wants to reveal the molecular mechanisms arising from all of the DNA differences in order to understand the biological processes taking place in human body. Figure 1 schematically illustrates harmless and disease-causing DNA differences, where different individuals carrying different DNA mutations are different in either their physical appearance (tall and short) or a predisposition to diseases (healthy and sick).

This paper outlines the progress made in human genome sequencing and detection of DNA differences as the necessary first step for personalized diagnostics. It is followed by reviewing the advances made in methods for discriminating disease-causing and harmless mutations. Simply predicting
that a given DNA defect is disease-causing is not enough for an effective personalized treatment, and thus the paper proceeds to review the approaches and techniques for predicting the molecular mechanism of disease-causing mutations for the needs of personalized diagnostics, pharmacogenomics, and personalized medicine.

2. Progress Made in Genome Sequencing and Database Development

Personalized diagnostics and medicine cannot be developed without access to the genomic data of the patient, which, in turn, requires inexpensive and fast methods for individual genome sequencing [14, 15]. The progress made in developing methods and techniques for detecting genetic variations was recently outlined in several works [16–18]. These techniques are rapidly evolving and various companies promised or already achieved the goal of being able to sequence an entire genome for a price tag of $1,000 within a day [19, 20]. However, despite the success of whole-genome sequencing, it is now understood that the analysis of genomic variations with respect to disease susceptibility is a much more complicated process and requires significant efforts [21]. If the sequencing is attempting to target a particular disease—in the simplest case a monogenic disease—then the analysis of the variations within a particular gene is typically a quite doable task. However, if the question is broader and if one wants to investigate the whole genome of the currently healthy individual with the goal of predicting future disease-causing defects, the problem becomes enormously complicated. One ends up with thousands or millions of variants spread over many different genes and noncoding regions of the genome. Identifying which of these variants might be associated with disease predisposition is not a trivial task. To assist in solving this difficult challenge, the 1,000 genomes project was founded. This project is aimed at revealing human variations within the entire genome by using the whole-genome sequencing of the DNA of 1,000 volunteers from different backgrounds [22, 23]. This is intended to provide information on the most frequently observed DNA variations, and the data is now available on the internet [24]. One can assume that genetic variations identified within the 1,000 genomes project are not necessarily disease-causing, since the volunteers are healthy individuals. However, it should also be noted that some diseases have very late onsets and may not be manifested at this stage in the individual’s life.

The interpretation of the individual’s genomic data has another difficult component, CNVs. Although important and frequently associated with diseases, CNVs cannot be easily used to reveal the molecular effect of a disease. One can speculate that a larger number of copies of a given gene will automatically result in a greater expression level of this particular protein and that this is the cause of the disease. However, this mechanism will not be discussed in this paper, because understanding the effect requires a detailed knowledge of the biological reactions associated with the target protein and how the change in the concentration level of individual macromolecules will affect the cellular function.

The progress in this fast growing field prompted development of databases on various levels such as disease-oriented databases to databases storing all of the known human DNA variations (Table 1). The creation of such databases serves two very important purposes: providing benchmarks to test in silico predictions and providing template cases or patterns for the detection of disease-causing mutation(s). Perhaps the most popular disease-oriented database is McKusick’s Online Mendelian Inheritance in Man (OMIM) database [25, 26], which is a manually curated database of human genes and genetic disorders including genetic phenotypes [26]. Since its establishment in early 1960s, many researchers have contributed to various aspects of the OMIM database such as developing extra features resulting in OMIM derivatives as PhenOMIM (for phenotypic comparison) [27], OMiR (to reveal associations between OMIM diseases and microRNAs) [28], CSI-OMIM (assisting clinical synopsis search in OMIM) [29], CGMIM (for text-mining of cancer genes) [30], and many other applications. On the other end of the spectrum is the dbSNP [31, 32] database at National Center for Biotechnology Information (NCBI). As of December 2013, it contains more than 140 million single nucleotide polymorphisms (SNPs), and the rate of new submissions is constantly increasing. In terms of distinguishing between disease-causing and harmless mutations, typically, one would create a pseudo-harmless database by taking out all of the entries from dbSNP that are listed in OMIM [33]. The remaining SNPs can be considered neutral or harmless, although exceptions to this rule will always be found. It should be noted that many other databases exist as well, some that focus on a particular disease [34, 35], others that focus on nonsynonymous SNPs [36, 37] or regulatory SNPs [38], or one that focuses on a particular family of genes [39, 40].

Although the goal of this paper is not to provide comprehensive review of all existing human variation databases, the HapMap [41] project and database cannot be omitted. The goal of the HapMap project is to develop a map of the common patterns of human DNA sequence variation. It is intended to be used to provide information about genes and patterns causing natural differences among individuals.
### Table 1: Tools and methods for predicting effects of mutations.

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Name</th>
<th>URL</th>
<th>Description and advantages/disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1000 Genomes: A Deep Catalog of Human Genetic Variation</td>
<td><a href="http://www.1000genomes.org/">http://www.1000genomes.org/</a></td>
<td>Freely available, provides tools for various searches as within populations, allele frequency and linkage disequilibrium structure</td>
</tr>
<tr>
<td></td>
<td>HGMD</td>
<td><a href="http://www.hgmd.org/">http://www.hgmd.org/</a></td>
<td>Requires license, but is the most comprehensive database of human mutations and provides visualization tools and search engines</td>
</tr>
<tr>
<td><strong>Networking</strong></td>
<td>Cytoscape</td>
<td><a href="http://www.cytoscape.org/">http://www.cytoscape.org/</a></td>
<td>An open source software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles, and other state data</td>
</tr>
<tr>
<td></td>
<td>Network Portal</td>
<td><a href="http://networks.systemsbio.net/">http://networks.systemsbio.net/</a></td>
<td>Provides analysis and visualization tools for selected gene regulatory networks to aid researchers in biological discovery and hypothesis development</td>
</tr>
<tr>
<td></td>
<td>Network2Canvas</td>
<td><a href="http://www.maayanlab.net/N2C/#.UIfP4qJWAkI">http://www.maayanlab.net/N2C/#.UIfP4qJWAkI</a></td>
<td>Allows for generation of drug-drug similarity and functional association of gene canvases</td>
</tr>
<tr>
<td><strong>3D structure modeling</strong></td>
<td>Swiss Model</td>
<td><a href="http://swissmodel.expasy.org/">http://swissmodel.expasy.org/</a></td>
<td>Automated server allowing 3D structure of a protein to be predicted inputting the sequence of the target</td>
</tr>
<tr>
<td></td>
<td>I-TASSER</td>
<td><a href="http://zhanglab.ccmb.med.umich.edu/I-TASSER/">http://zhanglab.ccmb.med.umich.edu/I-TASSER/</a></td>
<td>Fully automated 3D structure predictor, especially useful for difficult targets (low sequence similarity to known structures)</td>
</tr>
<tr>
<td></td>
<td>Modeller</td>
<td><a href="http://salilab.org/modeller/">http://salilab.org/modeller/</a></td>
<td>Stand-alone 3D structure predictor. Can be installed locally and used for large-scale predictions</td>
</tr>
<tr>
<td><strong>Protein stability</strong></td>
<td>FoldX</td>
<td><a href="http://foldx.crg.es/">http://foldx.crg.es/</a></td>
<td>Fast assessment of the changes of the unfolding free energy caused by mutations using empirical formula</td>
</tr>
<tr>
<td></td>
<td>I-Mutant</td>
<td><a href="http://gpcr2.biocomp.unibo.it/%7Eemidio/I-Mutant/I-Mutant.htm">http://gpcr2.biocomp.unibo.it/%7Eemidio/I-Mutant/I-Mutant.htm</a></td>
<td>Very fast and accurate prediction of the changes of the folding free energy caused by mutations. The input can be just the sequence of the protein</td>
</tr>
<tr>
<td></td>
<td>ERIS</td>
<td><a href="http://dokhlab.unc.edu/tools/eris/">http://dokhlab.unc.edu/tools/eris/</a></td>
<td>Uses Medusa force field to make predictions of the change of the folding free energy caused by mutations</td>
</tr>
<tr>
<td><strong>Protein interactions</strong></td>
<td>BeAtMuSiC</td>
<td><a href="http://babylone.ulb.ac.be/beatmusic/">http://babylone.ulb.ac.be/beatmusic/</a></td>
<td>Evaluates change in binding affinity between proteins caused by single-site mutations in their sequence</td>
</tr>
<tr>
<td></td>
<td>ICM</td>
<td><a href="http://www.mosoft.com/gui/mutation-protein-binding.html">http://www.mosoft.com/gui/mutation-protein-binding.html</a></td>
<td>Computes the change in binding free energy of a protein complex upon mutation of a single residue. Requires license</td>
</tr>
<tr>
<td></td>
<td>Robetta</td>
<td><a href="http://robeta.bakerlab.org/alascansubmit.jsp">http://robeta.bakerlab.org/alascansubmit.jsp</a></td>
<td>Provides estimation for the binding free energy changes caused by mutations to Alanine</td>
</tr>
<tr>
<td>Purpose</td>
<td>Name</td>
<td>URL</td>
<td>Description and advantages/disadvantages</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>---------------</td>
<td>------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Subcellular localization and pH dependence</td>
<td>WoLF PSORT</td>
<td><a href="http://wolfpsort.org/">http://wolfpsort.org/</a></td>
<td>Predicts the subcellular localization of proteins based on their amino acid sequences. The predictions are based on both known sorting signal motifs and amino acid content</td>
</tr>
<tr>
<td></td>
<td>PredSL</td>
<td><a href="http://aias.biol.uoa.gr/PredSL/">http://aias.biol.uoa.gr/PredSL/</a></td>
<td>Stand-alone code utilizing neural networks, Markov chains, and HMMs for the prediction of the subcellular localization of proteins</td>
</tr>
<tr>
<td></td>
<td>UniLoc</td>
<td></td>
<td>This is a webserver using PSI-Blast to infer homology between query sequence and already annotated proteins</td>
</tr>
<tr>
<td>Macromolecular function</td>
<td>SNAP</td>
<td><a href="https://www.rostlab.org/services/snap/">https://www.rostlab.org/services/snap/</a></td>
<td>SNAP is a neural network-based method that uses in silico derived protein information (e.g., secondary structure, conservation, solvent accessibility, etc.) in order to make predictions regarding functionality of mutated proteins</td>
</tr>
<tr>
<td></td>
<td>MutationAssessor</td>
<td><a href="http://mutationassessor.org/">http://mutationassessor.org/</a></td>
<td>The server predicts the functional impact of amino acid substitutions in proteins</td>
</tr>
<tr>
<td></td>
<td>SIFT</td>
<td><a href="http://sift.jcvi.org/">http://sift.jcvi.org/</a></td>
<td>The prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences</td>
</tr>
</tbody>
</table>
Advances in Biology

[42–45] as well as the predisposition to diseases [46, 47], responses to drugs [48], and cell phenotype [49].

3. Progress Made in Developing Methods for Revealing the Molecular Mechanisms of Disease-Causing Missense Mutations

The progress made in developing approaches to reveal the molecular mechanism of disease-causing mutations is outlined in several reviews [50–52]. Here, we briefly summarize the major approaches and developments, focusing on those which allow not only for classification of mutations as disease-causing or harmless, but also for providing information on what the dominant molecular mechanism behind the mutation is (Table 1). The focus of this paper is utilizing structural information to deliver predictions; however, in principle, one can make reasonably specific predictions about the effect of mutations on the protein interaction network using sequence information only. Because of this, the discussion below begins with a networking analysis and other associated approaches and then outlines the progress made in the structural space, and finally it demonstrates how the structural information can be used to reveal the details of the effects of a mutation.

3.1. Progress Made in Networking. Every macromolecule participates in various interactions resulting in a complex network in the cell. Understanding the effects of mutations requires evaluating the corresponding effect on the entire network as discussed recently [53]. Such an analysis is crucial for understanding complex diseases, that is, diseases caused by mutations in several genes. The observation that the same disease can be caused by different mutations in different genes leads to the conclusion that the phenotype is caused by multiple modifications at the molecular level, perhaps by disrupting the same network components. Because of this, complex diseases are frequently referred to as diseases of pathways [53, 54]. Understanding the effect of genetic differences on the corresponding networks requires generating the network representation and mapping the differences onto it. Typically, this is done by generating a graph on which the genes are placed at the nodes (vertices) and the interactions are represented as the links (edges) between the nodes. Perhaps the most widely used resource for the visualization of such networks is Cytoscape [55–58], although many alternative solutions do exist [59–62]. The main challenge is to identify or predict which genetic mutations affect which interaction, in other words, how to best map the mutations onto the edges of the graph. In some limited cases, associating a particular mutation with a particular interaction can be done by extracting data from the literature, analyzing the 3D structure of the corresponding complex, performing docking and then analyzing the structure of the docked complex, or predicting residues that participate in the interaction (correlated mutation sites) [63, 64]. This is still one of the main bottlenecks for large-scale modeling. Even if the genetic defects can be successfully associated with the edges of the network and assuming that these mutations simply remove the corresponding edge (a very simplified assumption, since more frequently the mutations weaken [33, 65] or strengthen [66] molecular interactions, not completely abolishing them), the next question is to predict the effect of edge removal on the disease phenotype. Only if all these questions are properly addressed can a prediction be made as to what the molecular mechanism of given disease (utilizing networking approach) is and, in turn, be able to point out which molecular interactions are affected and how this affects the cellular function.

Another challenge is that human interactome is far from complete and there are many missing interactions which have not been discovered yet [67, 68]. In addition, there are also many interactions detected by high-throughput methods which may not be real physical interactions taking place in the cell [69, 70]. Combined with dynamic nature of interactome [71, 72], it is clear that significant work needs to be done to better understand how mutations affect the network and, in turn, how the changes in the interactome, local or global, are associated with the wild type function of the cell. In particular, it is important to take into account the redundancy in the human interactome to prioritize plausible genes involved in a disease [73].

3.2. Progress Made in Structural Genomic Consortiums and 3D Structure Predictions. Structural genomic consortiums are intended to promote development of methods, tools, and approaches to deliver the 3D structures of novel proteins [74–77]. Depending on the overall goal, the focus varies from determining the 3D structure of proteins found in the human genome, proteins of medical importance, or proteins from other genomes. In the process of selecting targets whose structures are to be experimentally determined, either by the means of X-ray crystallography or by NMR, researchers frequently pick up genes which represent large class of proteins with no 3D structure available [76, 78]. Such an approach is intended to result in an equally populated conformational space and to provide homologous 3D structures for a maximum number of protein sequences. With the ever-growing Protein Data Bank (PDB) [79, 80], which as of December 2013 has 96,596 experimentally determined macromolecular structures (including proteins, RNA, and DNA), the investigations focusing on a particular gene (protein) are frequently able to find either the 3D structure of the wild type protein or the structure of a close homolog in the PDB, with an unfortunate lack of membrane and scaffolding proteins. If the 3D structure of the target protein is not available, one should build a model using the most appropriate homolog(s).

There are many different approaches for 3D structure predictions, varying from homology-based to first-principle-based approaches [81–87]. While all these methods have strengths and weaknesses, from point of view of delivering high quality 3D models, including models for large proteins, the homology-based approaches are far superior to the rest. As summarized by Moult, there is a significant improvement in methods utilizing template-based approaches which can be seen comparing the results of tenth Critical Assessment of
Structure Prediction (CASP) experiments [88]. The resulting 3D models of individual macromolecules, especially if based on highly homologous template(s), are of a higher quality that allows for meaningful structural analysis [89, 90] and even for carrying out various energy calculations [91, 92].

At the same time, since practically every macromolecule is involved in various interactions including interactions with other macromolecules [93, 94], it is equally important to reveal the interacting partners and the structure of the corresponding protein complexes. Several databases summarize and provide details about such interactions [95–98], including the changes to the binding affinity caused by mutations [99]. While a significant amount of thermodynamics data exists, very few structures of macromolecular complexes are available (as compared with monomeric macromolecules) and therefore the structures have to be predicted in most cases [100–103]. The 3D structures are typically modeled via either homology-based methods [104–108] or docking [109–112]. The performance of these approaches is tested in the community-wide experiment on the Critical Assessment of Predicted Interactions (CAPRI) [113], and it was concluded that the performance of docking and scoring methods has remained quite robust but challenges still exist [113–116]. Either way, one needs either experimentally determined 3D structure or a high quality model of the corresponding macromolecular complex in order to carry out structural analysis and evaluate the various energy components [33, 65].

The above considerations are with respect to the wild type macromolecules, which from genetics perspective typically are referred to as dominant allele. It is quite unlikely to expect that the 3D structures of the minor alleles or rare/unique mutant macromolecules and the corresponding complexes will be experimentally determined independently. Instead, the mutant structures are built from the wild type structures by either side chain replacement [117–121] or insertion/deletion of a structural segment [122–124] and further structural relaxation [33, 65, 125–128].

3.3. Progress Made in Understanding the Details of Disease-Causing Mechanisms Utilizing Structural Information. Revealing the effect(s) of genetics differences on the wild type cellular function can be done either experimentally or in silico. It is quite unlikely that experimental approach will be applied for each individual case, due to the fact that experiments are time-consuming and may require a significant investment. Due to this, in silico approaches must be utilized. Since the goal is to reveal the details of the effect, not just the effect itself, one needs structural information. To reiterate once more, it should be clarified that, for example, a prediction that a given mutation destabilizes the corresponding protein, which can be done without structural information, is not sufficient for understanding the details of the effect. Instead, one has to be able to predict what the structural changes caused by the mutation(s) are and how these changes can be reduced or eliminated by small molecule stabilizers. Below, we review the progress made in several major directions such as predicting the effect on (Section 3.3.1) protein integrity [50], (Section 3.3.2) protein interactions [129], and (Section 3.3.3) protein subcellular localization and pH-dependent properties. We purposely focus on these directions because, in principle, these effects can be fixed with external stimuli, such as small molecules. Interested readers should be referred to several other review papers exploring different effects [5, 51, 52]. In the end, it is important to recognize that the most successful predictions are expected to be done addressing the effects above and simultaneously taking into account the specificity of the function of the corresponding target. However, frequently, the precise function or the details are unknown and have to be predicted. The necessity of revealing macromolecular function in terms of understanding the disease mechanism and the progress made in this direction are discussed in Section 3.3.4.

3.3.1. The Effect on Protein Integrity. The effect on protein integrity is typically assessed via predicting the changes of the folding free energy, conformational dynamics, and hydrogen bond networks [50]. With this in mind, one of the main obstacles in predicting if a given mutation is deleterious is the ambiguity of how large the deviation from native property of a given protein should be in order to be disease-causing. For example, some proteins are very stable having a large folding free energy and small changes caused by mutation(s) may not be deleterious. At the other end of the spectrum are intrinsically unstable proteins with a folding free energy of a few kcal/mol; for them, almost any change in the folding free energy is expected to be deleterious. In order to avoid this particular problem with respect to protein folding free energy, an approach was developed to mutate all native residues to the rest of ninety amino acids and to construct the mutability landscape to guide the selection of deleterious mutations [130]. Such an approach allows the decision to be made based on the energy landscape of each particular protein. Another investigation introduced quantities such as “tolerance” and “mutability” for mutation sites to indicate if the site itself can tolerate substitutions and also to detect if these substitutions are amino acid specific [131]. Various approaches exist to predict the changes of protein stability due to mutations [132–137]. The performance of such selected methods, including resources which do not utilize structural information, was reviewed in recent reports and it was indicated that the ability of the methods to deliver accurate predictions is quite limited [138] and better tools are required [139].

The above considerations focus mostly on protein folding free energy changes caused by mutations; however, of equal importance are the effects of the mutations on macromolecular dynamics and the details of hydrogen bonding, especially in the neighborhood of the active site. Alteration of the hydrogen bond network within the active site or other structural regions important for the biological reaction is typically always deleterious [126, 128, 140, 141]. Changes in macromolecular dynamics, especially for proteins whose function requires conformational changes, can cause diseases [66, 142–144]. These changes in the hydrogen bond pattern and conformational flexibility are typically predicted via standard
molecular dynamics or energy minimization simulations. Provided that the mutations do not cause drastic structural alterations, the existing molecular dynamics packages are quite successful in revealing these changes [50].

3.3.2. The Effect on Protein Interactions. Essential components of cellular machinery are protein-protein interactions. Any missense mutations, especially those at the protein binding sites, can affect the affinity and interaction rates as discussed in a recent review [129]. Currently, there are several structure-based approaches to predict the changes of the binding free energy due to missense mutations [132, 145–150]. These methods utilize the experimentally derived 3D structure of the corresponding protein-protein complex. If the structure of the complex is not available, the alternative is to dock the monomeric proteins, to predict the 3D structure of the complex, and then to evaluate the effect of the mutation on the binding affinity. The performance of such approaches to predict structural changes and changes in the binding affinity caused by mutations is reviewed in recent article [151] and it is concluded that significant improvement is needed to improve the performance.

Despite the fact that the existing methods are not particularly accurate to predict the exact changes of the binding free energy due to mutation, as can be seen from benchmarking tests against various databases of experimental data points [95, 97, 99], the predictions still can be used to evaluate the trend of the changes without being too concerned about the magnitude of the changes [33, 65, 66, 131]. In addition, the structures of the corresponding complexes, either experimentally available or modeled in silico, can be used for structural analysis to predict the effect of mutations [152, 153]. With this in mind, of particular interest is the inferred biomolecular interaction server (IBIS) at NIH/NCBI [154, 155]. Thus, one can use structural information to make a reasonable prediction about whether the mutation will be tolerated or not, that is, if the mutation will have drastic effect on the protein’s wild type interactions.

3.3.3. The Effect on Subcellular Localization and pH Dependence. Macromolecules carry out their function by sensing various environments and, particularly in the cell, are localized in different subcellular compartments or are trafficked across different compartments. Each subcellular compartment as well as different body organs has a specific characteristic pH as compiled in several reports [156–160]. Macromolecules must be delivered to the correct compartment in order to function properly and any mutation that changes the signal peptide will have a deleterious effect on the function [161–163]. In addition, any mutation that alters the pH-dependent properties, either the pH dependence of protein stability [156, 157] or the protein-protein interactions [156, 157, 160, 164, 165] (including the changes of protonation states [166, 167]), may be deleterious. Such an analysis is not easy to do since the decision about the effect must be taken into account along with the subcellular or organ characteristic pH where the wild type protein is supposed to function, which is information that is not typically available.

If the characteristic pH is known and the structures of the corresponding macromolecules and their complexes are available, then there are many in silico tools to predict the effect of mutations on the pH dependence of folding and interactions as recently reviewed [168]. Some of them predict the conformational changes and the changes of hydrogen bond patterns as well, providing additional information to be analyzed. The performance of the existing methods of pKa calculations is increasing the accuracy to much higher levels by reducing the overall error to less than 1 kcal/mol [169]; this range is frequently sufficient for analyzing the effect of mutations.

In terms of predicting the effect of mutations on the properties of the signaling peptide, one can assess the effect using various databases and servers of signaling peptides [170–172]. Although considerations must be made about the accessibility of the signaling peptide from the water phase, in most cases just the sequence information is needed to make the prediction.

3.3.4. The Macromolecular Function and Effects of Mutations. In the above paragraph, the macromolecular function was frequently mentioned and it was repeatedly said that the effects of mutations should be evaluated in terms of their effect on macromolecular function. However, there are still macromolecules in the human genome which are not annotated [173], even for those whose 3D structures were experimentally determined via Structural Genomics Initiatives, so termed orphan proteins [174, 175]. It is infeasible that these functionalities will be experimentally studied, and these proteins and RNAs should be annotated computationally [173, 176–179]. Having in mind the importance of developing in silico tools for functional annotation, recently, the first large-scale community-based critical assessment of protein function annotation (CAFA) experiment has begun [180]. The results from the first round are quite encouraging in terms of the fact that standard sequence-based approaches such as Blast are capable of detecting sequence similarity and thus of inferring function [181], but it was indicated that there is a need for improvement of currently available approaches [180]. The main challenges include the definition of protein function and evaluation of predictions to be independent of the dataset [181].

In conclusion of this section, it should be clarified that indeed the currently available methods for structure analysis and predictions, energy calculations, hydrogen bond network modeling, assessment of conformational dynamics, and functional annotations are not perfect and need improvement. Still, if applied together to study any particular macromolecule and its associated mutations, it typically delivers meaningful results as indicated by comparing with the experimental data of the relevant case studies [66, 92, 126–128, 131, 182, 183].

4. Personalized Diagnostics

Armed with the abovementioned tools, the ultimate goal is to be able to detect disease-causing DNA defects even before
the disease is clinically manifested [184, 185]; however, it is equally important to pinpoint the disease-causing effect [66, 92, 127, 128, 183] (Figure 2). The last case of investigations is essential for building a library of DNA defects associated with particular diseases, that is, database of genotypes causing particular disease [186]. The increasing number and size of such databases is essential for fast and precise diagnostics, since the only information required is the individual's genome. Once the individual genome is mapped onto the database of the diseases' genotypes, the prediction of the disease predisposal can be done instantly. Perhaps the best approach is to collect DNA samples from all individuals, especially individuals in their early life, make such a screening routine, and monitor the individual's health throughout their life.

While database of disease-causing genotypes is an extremely important health issue, there will always be new genotypes which cannot be detected by such an approach before the clinical manifestation of the disease occurs. To associate a new genotype with a particular disease and reveal the molecular mechanism behind it will require applying the approaches described above. Perhaps in some limited cases, the molecular mechanism and the disease association of these new disease-causing mutations will be revealed by the means of experimental techniques or in model organisms, and then they will be added to the appropriate genotype database. However, in the vast majority of the cases, the molecular mechanism will have to be revealed in silico. Essentially, one should be able to address the following hypothetical scenario and provide a diagnosis for a particular individual: given an individual's genome, the goal is to identify all the potentially disease-causing mutations by comparing them to the databases of disease-causing genotypes. Then, the rest of the individual's DNA differences (with respect to the “standard” human DNA) must be analyzed in silico and disease-causing mutations must be identified among the DNA differences causing natural differences in human population. However, the completion of such a task is not trivial, because not only the distinction between disease-causing and harmless mutations is difficult, but also, more importantly, the linkage between predicted disease-causing mutations and the disease is extremely challenging, especially with complex diseases. Still developing biomarkers to individualize cancer treatment by identifying cancer-associated genes that can differentiate one type of cancer from another will enable the use of highly tailored therapies [187]. The problem is slightly less complicated for monogenic diseases, since the disease is known to be caused by the malfunction of a particular gene (protein) and if the given mutation in this protein is predicted to be disease-causing, then, most probably, it is associated with the same monogenic disease. However, notable exceptions do exist, as, for example, missense mutations occurring in MECP2 gene and causing either Rett syndrome [188, 189], Huntington’s disease [190], or other disorders [191].

5. Pharmacogenomics

With ever-increasing amount of clinical data, it is now widely understood that different races [192], ethnicities [193, 194], genders [195, 196], age [197, 198] groups, and so forth respond differently to various medications (Figure 2). A drug which is quite efficient for the treatment of a particular disease for a group of people sharing the same or a similar genotype may not work well for another group of people belonging to a different genotype. This may result from different phenotypes
of the disease among these groups of people, but even if the phenotype is the same amongst the group members, still the efficacy of the drug may depend on the differences in the genotypes. A prominent example of differing drug responses is human cytochrome P450 [199]. One of the isoforms of human cytochrome P450, CYP2D6, is primarily responsible for metabolizing hydrocodone to hydromorphone, a typical drug treatment after surgery [200]. However, it was found that a variant of CYP2D6, the CYP2D6.17 common in African Americans, does not metabolize hydrocodone efficiently [201]. Having prior knowledge of such cases and even more importantly being able to predict the drug efficiency based on the patient's genome is crucial for successful treatment. If such information is readily available, then the prescription can be personalized by prescribing different dosages depending on the patient's genotype. Even further, frequently there are several drugs designed to treat certain diseases and the selection of the best drug for the treatment should be based on the patient's genotype as well. Currently, the data is very scarce [202–204] and much work must be done in order to make pharmacogenomics a more common practice.

6. Personalized/Precise Medicine

The culmination of the usefulness of the individual's genomic data resides in personalized medicine [205]. The basic concepts of personalized medicine, or sometimes called precision medicine, are outlined in a recent article [206]. Essentially, it is a combination or a joint venture of personalized diagnostics, pharmacogenomics, and personalized preventive care [207–209] (Figure 2). Since personalized diagnostics and pharmacogenomics were already discussed above, the main focus here is the personalized preventive care. Ignoring ethical issues associated with providing individuals with predictions about their long term health [210], an early preventive treatment for plausible disease would have enormous effect on society and the individuals themselves. Perhaps, preventive care can be divided into several categories: (a) preventive care for conditional diseases; (b) preventive care for development diseases; and (c) preventive care for an individual's lifetime.

The most easily addressable preventive care is the care for individuals who may develop a disease which depends on certain (environmental) conditions. Obviously, avoiding these conditions will dramatically decrease the disease risk. For example, Chronic Beryllium disease is a disorder found in some individuals being exposed to Beryllium [211] in addition to having a particular genotype. If every individual applying for a job in Beryllium rich environment is genotyped and individuals possessing the risk genotype are notified of this risk and potential dangers, then, this would be the best preventive care for people susceptible to Chronic Beryllium disease. Other examples are the cases of individuals predisposed to lung or skin cancers [212, 213]. These individuals should avoid smoking or exposure to intense ultraviolet light, respectively. The list of examples can be extended to many other cases, but the message is that clear identification of individuals predisposed to diseases whose development depends on certain conditions would greatly decrease their reliance on medical treatment later on in life. In addition, in mental disorders the susceptibility profile of each individual depends on the psychosocial environment and this should be taken into account in delivering the prognosis [214].

Developmental diseases are typically quite severe and even if the patient survives, the effects are often permanent. Another important distinction between developmental diseases and other diseases is that once they are clinically manifested, it is typically too late for treatment. Due to the severity of these diseases, predicting an individual's genetic predispositions must be done at a very early stage in their development and the appropriate treatment must be administered [215].

Finally, there are many diseases and conditions which require a lifetime of care [216]. It is desirable that such cases are detected before the patient becomes sick. However, the preventive care in such a case, when the disease is still not manifested, will require quite different (from current) thinking from both the patient and the primary physician [217]. It may require decisions which will be difficult to justify without presence of the disease and in some cases may result in the wrong treatment. The straightforward solution is to avoid radical interventions but to subject these high risk patients to constant monitoring and frequent examinations.

7. Concluding Remarks

This paper attempts to outline the current development taking place in several rapidly evolving disciplines: personalized diagnostics, pharmacogenomics, and personalized medicine, and also how structural and conventional biology and in silico biophysics are embedded in these efforts. It is quite likely that individual genotyping will become a standard test, similar to currently used blood test, and the decisions about individual's health will be based on the corresponding genotype. The decisions about their health for either personalized preventive care or personalized treatment will be still individualized but not to the extent that each person will receive an individualized drug; rather, both the preventive care and drug prescription will be grouped into categories depending on common genotypes and phenotypes. With this in mind, structural and functional genomics along with better computational approaches will play crucial roles in the development of these methods.

However, many challenges still exist in fully utilizing genomic data to guide personalized medicine and pharmacogenomics [218]. Recent completion of the 1000 genomes pilot project [219] revealed that most individuals carry 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders [220]. In addition to this observation, it is known that the severity of a disease depends on many factors, and, for individual carrying the same disease-causing mutation(s), the manifestation can be quite different. At the same time, it was pointed out that disease-associated variants differ radically from variants observed in the 1000 genomes project dataset [221], providing a hope that, despite the natural complexity, the genetic
information will be used to provide better diagnostics and treatment.

It should be pointed out that it is clear that personalized medicine and pharmacogenomics will never be totally “personal.” The time and the effort to bring scientific discovery to the clinic, including the time for clinical trials, are prohibitively large and cannot be done on an individual basis. Instead, the causes of the diseases should be generalized into classes and specific, “individualized” treatment should be offered depending on individual’s DNA defect falling into a specific class for which particular treatment does exist.

Conflict of Interests
The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment
The work was supported by an institutional grant from Clemson University, the office of the Provost.

References


Advances in Biology


[128] L. Boccuto, K. Aoki, H. Flanagan-Steet et al., "A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt...


Submit your manuscripts at
http://www.hindawi.com