[Frontiers in Bioscience 13, 3391-3407, May 1, 2008]

From protein-disease associations to disease informatics

Mehmet M. Dalkilic, James C. Costello, Wyatt T. Clark, Predrag Radivojac

School of Informatics, Indiana University, Bloomington, IN 47408, U.S.A.

TABLE OF CONTENTS

1. ABSTRACT

Advancements in high-throughput technology and computational power have brought about significant progress in our understanding of cellular processes, including an increased appreciation of the intricacies of disease. The computational biology community has made strides in characterizing human disease and implementing algorithms that will be used in translational medicine. Despite this progress, most of the identified biomarkers and proposed methodologies have still not achieved the sensitivity and specificity to be effectively used, for example, in population screening against various diseases. Here we review the current progress in computational methodology developed to exploit major high-throughput experimental platforms towards improved understanding of disease, and argue that an integrated model for biomarker discovery, predictive medicine and treatment is likely to be data-driven and personalized. In such an approach, major data collection is yet to be done and comprehensive computational models are yet to be developed.

2. INTRODUCTION

In the past three decades, a bioinformatics community has arisen as a result of inexpensive, powerful computers, the Internet and advancements in high-throughput, genome-wide technologies. Like most nascent areas, the boundaries of bioinformatics are not clearly defined and there are perspectives to reconcile. However, there does exist a recognizable direction of research, including a set of *grand challenges* for the community (1, 2). In addition, there are many instances where bioinformatics has significantly contributed to filling in the missing pieces of the puzzle of life or has provoked researchers to rethink existing biological knowledge (3).

A considerable amount of the work in bioinformatics is directed towards understanding how to both deal with and interpret massive amounts of biological data. Much of this work is also devoted to the development of systematic and theoretically well-founded approaches to problems that allow sharing of results. For example, the

need to use structured vocabularies in order to maximize the advantages of powerful computational approaches has stimulated the widely adopted genomics standards of the Gene Ontology (GO) (4). In addition, the statistical sciences have been reinvigorated by the research in life sciences, while biology has benefitted from the rigorous analysis of the data sets far larger than biologists had previously acquired. Sharing and development of techniques across experimentally distinct data leads to a community of researchers, albeit composed of different kinds of scientists, who leverage results from areas other than their own to solve biological problems.

Although the studies of disease sit squarely within biology, and the challenge of translational medicine has been articulated (2), an analogous community to bioinformatics seems not to have been formed yet. Also, though there are certainly successes in computational approaches to studies of disease, or what we will refer to as *disease informatics*, there is a lack of cross-fertilization among groups approaching similar problems from different perspectives. As a result, approaches to anticipating disease, prognostics, treatments and drug design have yet to meet expectations. In fact, the methodologies developed to address disease-mostly based on statistical and physical principles-are typically exploiting individual technological advancements such as gene expression arrays, protein arrays, tandem mass spectrometry, *etc*. As these methods have matured from development to application, disciplines like systems biology are being developed to approach the problem from a different perspective and integrate diverse data types (5). It remains an open problem, however, how to appropriately address disease and integrate individual successes of disease informatics in a way that is most useful for translational approaches and, ultimately, to form a community much like bioinformatics.

In this paper we first provide an extensive survey of computational approaches in the studies of disease, and then conclude with the proposal of a high-level quantitative model of disease informatics.

3. DISEASE INFORMATICS

In this Section, we summarize the most recent efforts in computational approaches to study disease. Several studies have recently addressed various aspects of disease informatics, providing a case-based structural perspective (6), analysis of algorithms for gene prioritization (7), analysis of protein interactions and disease (8), high-throughput phenomics (9) and complex networks (10). We, however, aim to take a broader approach that takes the most recent results from several communities and subsequently propose a model in which such individual approaches can be integrated. We start by addressing the importance in organization of disease classification and proceed to critically discuss a variety of areas currently developing in the bioinformatics community: the importance of model organisms for disease informatics, the underlying characteristics of disease-associated genes, structural and protein folding approaches to studying disease, transcriptomics, proteomics and metabolomics. All of the individual areas provide quantifiable data on the state of a cell or group of cells and it is our conjecture that these individual components can systematically be brought together to better inform the study of disease. Therefore, the topical discussions in the following subsections lay the groundwork for a proposed model for disease informatics, which will be introduced in Section 4.

3.1. Controlled vocabulary for disease naming and hierarchical organization of disease terms

Although the importance of disease classification was understood as early as in the Hellenistic world, where symptoms were grouped together to be treated similarly, the first attempt at a standardized nomenclature originated in the mid-18th century from the need to classify and statistically process the causes of death (11). A century later, many conditions were observed to be non-terminal, giving rise to the first international attempts at statistically characterizing disease that resulted in the first International Classification of Disease (ICD). This list, currently available as revision 10 (ICD-10), is maintained by the World Health Organization (12). In the last quarter of the 20th century, however, a requirement of not only controlled vocabulary, but also relations between terms gained importance in order to facilitate computing. Systematized Nomenclature for Medicine, Clinical Terms (SNOMED CT) (13) is an ontology that has been developed to remove the semantic differences between terms in a medical setting. With over 310,000 terms organized into 19 hierarchies, such as clinical findings, procedure and body structure, SNOMED CT offers an expansive and flexible means to describe and organize clinical observations. The biggest use of SNOMED CT is in patient Electronic Medical Records that are being implemented in health care systems around the world. Another example of structured terminology is the Unified Medical Language System (UMLS) developed by the National Library of Medicine, which addressed the need to unify divergent naming and provides a means for computer processing of disease (14). The UMLS is a comprehensive system consisting of a more than 2.5 million terms for about 900,000 biomedical concepts (15) organized through the Metathesuarus (which contains the entire SNOMED CT), Semantic Network and a SPECIALIST Lexicon, but it is significantly broader than the classification of disease. Researchers at Northwestern University have created the Disease Ontology (DO), a controlled vocabulary based on the subset of UMLS terms and the ICD. Similar to the use of the GO in the classification of biological process, molecular function and cellular component of gene products, the DO[1] hierarchically organizes disease at different levels of specificity into a directed acyclic graph. Currently, DO ver. 2.1 contains 14,647 terms classified into no more than 15 hierarchical levels, starting with the root node Disease, and providing the appropriate structure for automated analysis of phenotypic function. Finally, automated approaches to phenotypic classification have also been developed. PhenoGO (16) combines the natural language processing system, BioMedLEE (17), MeSH terms and the PhenOS system (18) to automatically assign a phenotypic context to genes and GO term annotations, where the context of a phenotype can be cell type, disease, or tissue to name a few.

Table 1. Summary of differences between disease-associated genes and non-disease associated genes

Differences between disease-associated and non-disease associated genes
• disease-associated genes are on average longer than non-disease associated genes
• disease-associated genes are more likely to have homologs in distant species, but less likely to have close paralogs than non-disease associated genes
• disease-associated genes have more interacting partners on average than non-disease associated genes, but fewer than essential genes (approximated by using housekeeping genes)
• certain biological functions are overrepresented or underrepresented in disease-associated genes
• disease-associated genes have more exons and greater total exon length than non-disease genes, but they have similar total intron length and 3' and 5' UTR region length
• disease-associated genes do not have distinct K_a/K_s values as compared to the remaining genes (except slightly for essential genes)
• disease-associated genes are less conserved than the essential genes, but similarly conserved as the remaining non-disease genes
• genes associated with the same diseases tend to interact more frequently, are co-expressed in the same tissue and tend to share GO terms
• disease-associated genes that correspond to certain GO terms have similar modes of inheritance (dominant vs. recessive)
• disease-associated genes generally have higher expression levels than non-disease genes but are expressed in a narrower range of tissues

3.2. Human disease addressed though animal models and cell lines

Model organisms have long been used to study the underlying mechanisms of biology, including disease (19-22). The sequencing, assembly and annotation of the yeast, worm, fly and mouse genomes, to name only a few, have now allowed researches to quantify the number of genes and molecular pathways related to human disease that can be studied in these organisms. For example, by utilizing the Inparanoid algorithm for homolog identification, it has been reported that approximately 57% of human disease genes have homologs in mouse, 51% in fly, 37% in worm and 14% in yeast (23). These data, along with homologous genes from several other organisms are stored in the online database, Orthodisease (24). In addition, it has been shown that disease-associated genes share more homologs with model organisms than genes currently not associated with disease (25, 26).

Many high throughput data sets from model organisms can be used to inform human molecular relationships (27-31). The concept of an "interolog" and "regulog" have arisen from the need to infer functional relationships across organisms (32-34). An interolog refers to an interacting protein pair in one organism that has homologus proteins in another organism that also interact, while a regulog refers to a homologous protein pair in two different organisms that share a homologous regulatory mechanism. Both of these concepts have been used to infer functional relationships in human, which is evident in the Interologous Interaction Database (I2D) (formerly the Online Predicted Human Interactions Database or OPHID) (35, 36). This database contains experimentally tested human protein interactions, interologous protein interactions and predicted protein interactions.

Although model organisms can be extremely helpful in understanding biological mechanisms, they have the fundamental flaw of not being human, and therefore cannot fully represent human cellular behavior. For example, certain disease-associated mutations in humans correspond to the wild type genotype in various model organisms (37, 38). Given that genetic manipulation cannot be performed on humans, the closest cellular representation of manipulatable human cells is derived cell lines. There are thousands of human cell lines that cover a multitude of tissue types and tissue conditions. Human cell lines can be exposed to the same, if not more, experimental conditions as model organisms. For example, microarray studies have been carried out with the NCI-60 lines to systematically identify differential gene expression between all of the 60 lines (39).

3.3. Disease genes are distinct, but why?

The first studies investigating the statistical, functional and evolutionary properties of disease-associated genes appeared in the 1990s, during the time of the Human Genome Project. In 1997, Mushegian *et al.* (25) carried out a systematic analysis of disease-associated genes by looking for their homologs in several model organisms and by analyzing their function and length distribution, but the main progress in understanding properties of disease-associated genes came after the completion of the human genome (40, 41). Although Mushegian *et al.* noticed that half of the known disease genes in the mid 1990s were involved in cell signaling, Jimenez-Sanchez *et al.* (42) found a number of other functional properties that differed between classes of disease genes (*e.g.* mode of inheritance, age at onset, frequency of disease), while Karlin *et al.* (43) proposed that disease-associated genes have distinct sequence properties when compared to non-disease genes, especially with respect to sequence runs. This analysis was broadened by Lopez-Bigas and Ouzounis (26) who also found that disease-associated genes are on average longer and expected to contain more homologs with distant species in the evolutionary tree, but that non-disease genes are expected to have more paralogs in the human genome. The evolutionary properties of disease genes were additionally analyzed by other groups (44-47) with an important result coming from Tu *et al.*, who found that disease and non-disease genes have a similar ratio of non-synonymous (K_a) and synonymous (K_s) substitutions, but that human essential genes (approximated by the housekeeping genes) have a slightly lower K_a/K_s (47). Similarly, exploiting the availability of protein-protein interaction networks, Tu *et al.* found that disease genes generally have higher connectivity than non-disease genes and that both groups have lower connectivity than the housekeeping genes. Goh *et al.* (48) further analyzed connectivity and other properties of disease, non-disease and housekeeping proteins and constructed a human disease network. Note that in addition to the comprehensive analysis of all disease genes, studies have investigated properties of proteins involved in specific diseases, for example, cardiovascular disease (49), or cancer (50-53). A summary of differences between disease and non-disease genes is summarized in Table 1.

Based on various properties discussed above, Lopez-Bigas and Ouzounis (26) developed a decision tree for identifying disease genes with surprisingly high recall/precision balance of 70%/67%. Adie *et al.* (54) improved on this method by employing a machine learning classifier based on a variety of genomic and evolutionary features, namely coding sequence length, presence of signal peptides, the number of exons, evolutionary conservation, presence and closeness of paralogs in the human genome. Xu and Li (55) developed a k-nearest neighbor algorithm to discriminate between disease and non-disease associated genes based on five properties calculated from the protein-protein interaction network: connectivity of the node, the fraction of links to the disease genes, the fraction of the disease associated neighbors to the node, average distance to the disease genes and positive topology coefficient. The accuracy achieved by their algorithm reached 74% recall and 75% precision on a literature curated set of gene-disease associations. The major limitation of these algorithms, however, is that they are exploiting only general properties of disease-associated genes and are not suited to prioritizing genes for any disease in particular.

While these studies are informative, larger questions remain: why are disease-associated genes different than the remaining human genes? Are there biological reasons or is this a statistical artifact of the diseases studied thus far? Why are all genes not disease-associated? Perhaps some of the answers can be inferred from taking an evolutionary, functional and statistical perspective. For example, strong purifying selection would be expected to act on genes that are critically involved in fetal survival or reproductive capacity before the reproductive age, whereas such pressure would be significantly lower on the genes involved in late-onset diseases. A functional perspective suggests that some functions of non-disease genes could be sufficiently well performed by their close paralogs for an extended period of time. Finally, a simple probabilistic calculation, assuming a uniform mutation rate, suggests that current disease genes, being generally longer than the non-disease genes, have simply had higher chance of acquiring disease-causing mutations over time. On the other hand, every gene has a molecular function and it should not be excluded that a disruption of every function would have phenotypic consequences. Since it seems unlikely that this difference is attributable to the sampling bias of disease selection, questions about disease-associated genes will remain open and further research is necessary to provide more definitive answers.

3.4. Algorithms to predicting gene-disease associations

With the accumulation of large amounts and multiple types of experimental data, prediction of gene-phenotype associations has emerged as a very productive subfield with great importance for the understanding of human disease. Given a particular set of human phenotypes (typically diseases) D, a set of human genes G and evidence E, these methods attempt to find whether gene $g \in G$ is associated with phenotype $d \in D$. Note that evidence E can be gene-disease associations obtained through genetic studies, experimentally determined protein-protein interactions (PPI), microarray data, but also gene function, protein sequence, biomedical literature, predicted PPI, *etc*.

A crucial piece of evidence used for the prediction of gene-disease associations is provided by the statistical genetics community, through linkage analysis and association studies (56, 57). However, due to the limitations of these approaches caused by genetic heterogeneity, small and biased population samples, as well as low penetrance rates (58), other sequence and physical data are useful in inferring gene-disease associations. A wide range of methods developed to infer these associations typically use statistical, machine learning or heuristic approaches for gene prioritization. While there are studies based solely on PPI data of humans and other species (59-62), novel disease candidates can be best inferred by combining multiple lines of evidence. Perez-Iratxeta *et al.* (63, 64) and Seki and Mostafa (65) achieved this by calculating gene-disease associations by linking phenotype to GO terms via text mining Medline articles, whereas Hristovski *et al.* (66) used association rules. Another tool, POCUS, calculates the probability that different loci share observed InterPro domains and GO terms by chance (67). Similarly, the same types of evidence were combined with sequence and evolutionary data through a decision tree by Adie *et al.* (54, 68). Gentrepid (69) presents heuristic prioritization based on PPI data and domain sharing. The method by Freudenberg and Propping (70) uses phenotypic data from OMIM to cluster diseases and then scores each gene-disease relationship (g, d) proportional to the shared GO annotation between a query gene and disease clusters associated with d. TOM combines functional and microarray data to statistically score genes most similar to a set of seed genes (71), while Prioritizer further incorporates PPI data via a Bayesian approach (72). An evaluation on the contribution of multiple data types was carried out by Aerts *et al.* (73) and De Bie *et al.* (74) who used statistical and machine learning principles, respectively. Evidence from *D. melanogaster* has also been used to infer human-disease relationships by Costello *et al.* (75), through machine learning and utilizing sequence similarity, GO annotation, protein interactions and shared transcription factor binding as inputs. Finally, larger efforts have been made to associate genes with phenotypic concepts via data mining (76) and with a totality of environmental and phenotypic concepts via statistical analysis (77). For example, Butte and Kohane create a phenome-genome network by associating genes to UMLS concepts (77).

In addition to the approaches associating entire genes to phenotypic terms, Braun *et al.* (78) proposed a method that identifies regions of candidate genes that are most likely to contain disease-causing mutations. The method is based on the fact that disease causing mutations are more likely to occur in the most conserved sequence regions. Thus, instead of screening full genes in their entirety, the authors suggest screening a larger number of genes in regions that are more likely to contain disease mutations.

Clearly, gene-phenotype association studies have explored a large number of approaches and several software packages or web sites are publicly available (64, 68, 69, 71, 73). However, it should be noted that many obstacles need to be overcome before these algorithms can fully impact translational research. First, it is somewhat vague what a gene-disease association means. Some authors only investigate causative associations where mutations are known to be hereditary. For others, associations can involve downstream effects on protein molecular function, such as genes affected by somatic mutations or epigenetic causes, because such genes can be good drug targets. Most problems seem to come from the fact that data are noisy and sparse, that disease names are stored in a variety of formats and that the community has not yet widely accepted any ontology of diseases. For example, PPI data is non-randomly incomplete (it is currently difficult to obtain interactions for membrane proteins) and might contain large number of false positives (79), while questions still remain about the biological reproducibility of microarray data (80). Similarly, in gene association studies, it is not uncommon that gene-disease associations are incorrect due to ascertainment bias and small sample size (*e.g.* in the case of obesity-related genes (81)).

3.5. Protein sequence, structure, function and folding in understanding disease

Currently, there are close to six million validated single nucleotide polymorphisms (SNPs), of which more than 50,000 are associated with various human diseases (82). Their functional effects include altered gene transcription/regulation, RNA decay, protein translation, signal transduction or changed structural integrity of cells (82). However, only for a small fraction of SNPs have a known molecular basis of disease (83, 84). Most research attention in the bioinformatics community has been devoted to investigating structural features of non-synonymous coding SNPs (nsSNPs) (85-89) and their connections with disease (90-94). For example, it is estimated that about 20-30% of nsSNPs stored in dbSNP (95) could affect protein function (96, 97) and that more than 80% are located in structural pockets or voids (91). Recent analysis of human disease proteins shows that a large number of nsSNPs appear in patches on the protein surface suggesting that the mutations might be directly affecting protein interaction sites (93) in addition to the known effects on structural stability (92, 98). Surface accessibility and evolutionary conservation have been recognized as the most useful features in the predictors of deleterious mutations (87, 88).

Though there is little doubt that protein structure is one of the keys to understanding the molecular basis of disease, there exist only a handful of success stories. Classic examples include the E6V mutation in the β-subunit of hemoglobin (HBB), which causes aggregation via interacting with F85 and L88 of another identical molecule. This event is known to cause excessive formation of amyloid fibrils ultimately leading to abnormally shaped erythrocytes and development of sickle cell anemia predominantly in homozygous individuals. Another prominent example involves interaction between Mdm2 and p53, where overexpression of Mdm2, a negative regulator and E3 ubiquitination ligase for p53, causes inhibition of p53 and leads to several forms of cancer (99). The structural basis of this interaction has recently been solved, involving molecular recognition fragments (MoRFs) as an important structural motif in signaling. In addition, understanding of these interactions resulted in novel concepts of drug design. MoRFs are short, loosely structured sequence fragments located within intrinsically disordered regions which often serve as protein-protein interaction sites (100-102). Computational methods have been developed for prediction of MoRFs (103-105). In accordance with this, most methods developed for prediction of protein global and residue-based function can be used towards better understanding and, in the end, treatment of disease. Some examples include prediction of protein molecular function, prediction of protein interaction sites and partners, protein interface residues and hot spots.

It is well-understood that after synthesis in a ribosome, the fate of a protein can be determined by a variety of effects in a cell, causing it to misfold, to be prematurely degraded, to aggregate or to form amyloid fibrils. It has been shown that proteins that show strong propensity to fibrillate can be either structured or unstructured, with the structured class covering all major groups (α, β, α/β) (106, 107). Uversky and Fink analyzed several classes of amyloid-prone proteins from structural perspective showing that structured proteins require partial local unfolding in order to fibrillate, while intrinsically disordered proteins require partial folding (106). Many proteins which are known to have strong propensity to form amyloid fibrils are still not linked to any disease, for example, myosin and prothymosin. Fernandez *et al.* proposed a concept of dehydrons, *i.e.* underwrapped (unprotected) backbone hydrogen bonds, in monomers and suggested that such structurally-encoded sites are important for acquiring new protein interactions (108-110). Interestingly, the analysis of multiple protein structures has provided links between dehydrons and protein amyloidogenic propensity (111).

Another direction in exploring associations between biological macromolecules and disease is through the analysis of folding pathways. It has been estimated that about 50% of human diseases are caused by protein misfolding events (112), which often lead to loss/gain of function in numerous *protein folding diseases* such as cancer, osteogenesis imperfecta, sickle cell anaemia, Alzheimer's, Parkinson's, and Hungtington's disease, to name a few. (113). The last 30 years have brought dramatic progress in the simulations of protein folding where μs simulations at fs resolution of smaller macromolecules are feasible (114, 115). However, there is yet to be enough progress in studies of larger molecules, especially supramolecular complexes or modeling solvents and crowded conditions of the cell (115, 116). Despite this, there is sizeable amount of work in studies of the mechanisms of disease, including simulations of influenza virus activity (117-119), prion misfolding (120) and aggregation of β-amyloid peptides in Alzheimer's disease

(121). There is agreement that both increase in computational power and better modeling of multi-body force fields are important for the further progress in molecular dynamics simulations.

3.6. Gene expression, genomic variation and disease

Gene expression measurements on a genome-scale, representing the transcriptome, have been accomplished through the technological advancement of microarrays. Introduced in 1995 by Schena *et al.* (122), microarray experiments have rapidly grown in number and have been used to address a multitude of questions across hundreds of organisms and cell lines. Related to human disease, microarrays have been used to define cancer-specific expression profiles (123, 124), association of SNPs within complex disease (125, 126) and shown copy number polymorphism in the study of cancer (127, 128). However, as noted by Miklos and Maleszka (129), there are limitations to what microarray data can tell us in the context of complex disease. Using schizophrenia as an example, the authors point out that differentially expressed genes may not correlate well across microarray platforms, and genes known to be involved in schizophrenia may not correlate well with genes identified as "interesting" in microarrary data. These discrepancies show the complications in separating the causal relationships from noise within microarray data. The authors also point out the differing results can be derived from different bioinformatics methods and call for integrating the results from these methods to assign a "clinical relevance" score to genes with consistent identification. Notwithstanding this cautionary note, the massive amounts of data that microarrays generate have led researches to explore different data mining techniques and meta-analysis methods to extract hidden relationships, which have shed light on gene expression associations in disease.

Golub *et al.* (123) were among the first to use microarray data to computationally define tumor subtypes and classify tumor sample expression patterns. The authors used clustering and a weighted scoring scheme to classify acute lymphoblastic leukemia and acute myeloid leukemia simply based on the expression profiles of the 50 genes most differentially expressed between the two conditions. A similar approach taken by Alizadeh *et al.* (130), applied hierarchical clustering to gene expression profiles of diffuse large B-cell lymphoma to identify expression patterns reflecting the stages of tumor differentiation. van't Veer *et al.* (131) also used hierarchical clustering with supervised methods to classify expression profiles of breast cancer patients based on time to metastases. This approach is significant since the standard medical treatment through chemotherapy or hormonal treatment is unnecessary in 70-80% of the patients receiving treatment, thus, this method can provide valuable information to clinicians when determining treatment type. Furey *et al.* (132) apply the classification propensities of support vector machines to distinguish between ovarian cancer tissue, healthy ovarian tissue and other healthy tissues, which results in perfect classification of the tissue types, and is even sensitive enough to identify a mislabeled tissue. Many methods exploring expression profiles examine individual gene profiles, but there still remains the question of how expression in particular classes of genes are affected. Several statistical methods, known as gene set enrichment (133-135), have been proposed to look at the overall expression of predefined sets of genes, such as genes annotated with the same GO term. Mootha, *et al.* (133) used the gene set enrichment approach to show that the genes involved in oxidative phosphorylation were showing group-wise expression bias in diabetes microarrays. This is significant, because none of the individual genes annotated with the GO term (oxidative phosphorylation) showed differential expression suggesting a molecular mechanism that was not represented by a single gene.

Clinical experiments employing microarrays as an assay for cancer gene expression have grown in number, and continue to do so. There have been several studies performed by independent institutions that explore the same cancer type, for example prostate cancer (136-139). These cross-laboratory and often cross-platform results provide a perfect scenario to employ meta-analytic techniques to discover the strength of relationships that exist within compendia of data sets. Meta-analysis, which combines data using statistical methods, has been applied by Rhodes *et al.* (124) to the aforementioned prostate cancer datasets to cross-validate individual results and find the expression profiles that were consistent among all prostate cancer studies. The identified sets of genes across each of the data sets were subsequently found to have common metabolic function. Spurred on by this methodology is the online resource, Oncomine (140), which is a repository that contains standardized, normalized and analyzed data from over 20,000 cancer microarray hybridizations. Not limited to cancer, gene expression assays and similar computation analysis have been performed in many other human diseases, including obesity and fatty liver disease (141), diabetes (142) and Lyme disease (143) to name only a few.

Besides the detection of variation in mRNA levels in diseased samples as compared to healthy samples, microarrays have also been adapted to test for copy number variation through comparative genomic hybridizations (CGH). This application of microarray technology has been used to show copy number aberrations in tumors (127) and also in genetic diseases (144). As an example, Vissers *et al.* (145) performed CGH analysis of patients suffering from CHARGE syndrome and identified a deletion on 8q12. Up to 15% (146) of monogenic diseases are related to deletions or insertions, thus the authors propose using the CGH platform as a tool to identify candidate genes involved in monogenic diseases. Since microarray-based assays have been shown to provide information on copy number variation and gene expression, Pollack *et al.* (128) combined data from both gene expression and CGH breast cancer studies to statistically show copy number variation has a large affect on global gene expression. As an attempt to identify literature using CGH for profiling malignancies, Progenetix[2] has compiled a compressive list of close to 1,000 publications, which demonstrates the impact of CGH.

Large-scale genotyping through the use of SNP-based microarrays is another strength of the microarray

platform. As an example, Matsuzaki *et al.* (125) simultaneously genotyped samples for over 100,000 SNPs with very high accuracy and reproducibility. The data from SNP genotyping technologies provide information that can be used in association studies, where variation in genetic markers (SNPs in this case) is associated with a particular phenotype (disease in this case). For example, Teh *et al.* (147) used SNP arrays to identify loss of heterozygosity on several chromosomes that are highly correlated to basal cell carcinoma. Another use of SNP array technology is exemplified by LaFramboise *et al.* (148), who use SNP array data to computationally quantify allele specific copy number and parent chromosome specific copy number through an expectation-maximization based method. Their results from the analysis of 100 lung cancer patients reveal that duplications are most often linked to only one of the parental chromosomes, which can be used as genetic markers.

The complications pointed out by Miklos and Maleszka (129) are at the heart of integrative studies, which leverage different data types to mine a rich set of information often hidden within microarray data. By integrating across data types, such as protein interaction and microarray gene expression, the strengths of each experimental technique can be leveraged. Bandyopadhyay *et al.* (149) incorporated network algorithms to identify transcriptionally active protein-protein relationships related to the reactivation of suppressed HIV in a host genome. The authors identify a number of genes and biological processes with their methodology. As another illustrative example, Valdivia-Granda *et al.* (150) integrate molecular annotation information, protein interaction data and microarray gene expression data to identify early infection biomarkers for smallpox. The authors find early infection biomarkers that are unique to smallpox and these markers represent a multitude of biological processes.

3.7. Proteomics, metabolomics and disease

Proteomics is a field whose goal is to identify and quantify proteins present in a sample mixture (151-153). It holds the promise of directly quantifying expressed proteins and/or determining the fraction of protein copies that are post-translationally modified and thus contribute in the early diagnosis of disease and biomarker discovery (154). However, vastly different abundances of different proteins and possibly small percentages of post-translationally modified copies pose significant challenges for current technology. For example, the quantity of proteins in human plasma is known to differ by 10-12 orders of magnitude and it is common that cancer biomarkers are present in ng/ml concentrations (155). Thus, despite being an older area of study than transcriptomics, proteomics is still in the early stages of development with an open challenge of what experimental platforms and computational techniques are best suited to it. Here we focus only on mass-spectrometry (MS) based proteomics techniques due to their high-throughput and increasing sensitivity. Other proteomics approaches are out of scope of this review and are reviewed elsewhere (154).

Most computational approaches in disease proteomics have been connected to surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) technology that is well suited for detecting intact proteins of moderately high molecular weight. The computational goal for SELDI-TOF data is often to discern between two groups of MS spectra, one corresponding to a group of patients with disease and another corresponding to the control group and then label the peaks that are statistically different (*i.e.* provide peak-to-protein assignments). SELDI-TOF has been used extensively in the biomarker discovery for several types cancers (*e.g.* prostate, breast, lung, ovarian (156, 157)), and a number of statistical models have been proposed and subsequently compared (158). However, SELDI technology, as well as mass spectrometry-based techniques in general, have been shown to be prone to error due to sample preparation and storage protocols which caused many biomarker discovery studies to be irreproducible (155, 157). In addition, peak labeling in SELDI-TOF platforms is an open challenge due to the presence of alternatively spliced isoforms and protein-modifications that cause a peak shift in the spectra. Tandem mass spectrometry (MS/MS) holds promise for a more accurate biomarker discovery. Novel experimental methods have been designed for sample fractionation and separation while computational approaches have been proposed for peptide identification (159-162), protein identification (163-166), protein quantification (167-169) and discovery and characterization of post-translational modifications (170-172). Further development of these methods will directly affect our ability to address biomarker discovery from MS/MS data.

Metabolomics attempts to study the large number of molecules that do not fall under the umbrella of proteomics and transcriptomics, typically referred to as the metabolome. The constituent parts of the metabolome are sometimes referred to as the set of molecules produced in metabolism (173), but a significant portion of metabolomics research studies lipid molecules (174-177) and other small molecules present in the body. The field, in its methodology of collecting and analyzing data, is closely related to proteomics. NMR spectroscopy and MS are the two most common experimental platforms used in metabolomics, while data analysis can also be divided into two areas, one being multivariate statistical methods for analyzing interactions between metabolites, and the other dedicated to identifying particular molecules (178, 179). Metabolomics also faces many of the same challenges as proteomics, especially regarding standardization of experimental procedures (179). Metabolite levels are also known to vary drastically both amongst and within individuals in some bodily fluids, especially urine. Although metabolomics has been an area of study for some time (180, 181), especially by the pharmaceutical industry (182), much research in metabolomics is hindered by the fact that much of the metabolome is uncharacterized (183). The creation of the Human Metabolomic Database (173) is a step towards developing a database that characterizes many these molecules

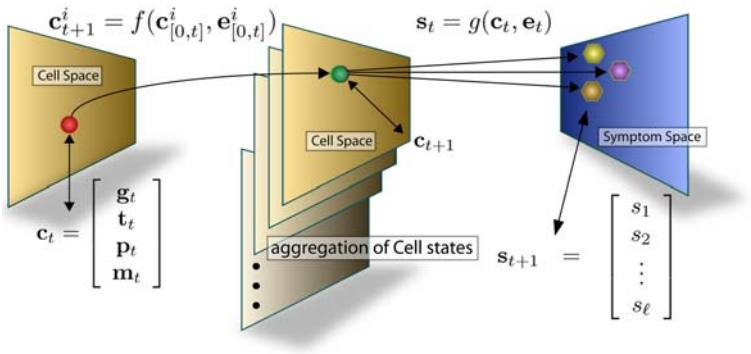

Figure 1. A stylized depiction of the model for the computational approaches to disease.

4. TOWARDS AN INTEGRATED COMPUTATIONAL MODEL FOR DISEASE

Despite the plethora of high-throughput technological advances and the success of many computational methods developed to understand disease, there is a growing need to work toward a model that allows for comprehensive and systematic successes in diagnostics, prognostics and treatment of the causes of disease. The benefits of such an integrated model would allow for (i) a data-driven, quantitative understanding of disease; (ii) disentangling symptoms from disease, (iii) taking advantage of the ability of technologies to query the states of single cells or small group of cells (*e.g.* flow cytometry or laser-capture microdissection) and (iv) more formalized and accurate notions of biomarkers.

The starting point for any model is to define the basal components that are going to be modeled. Therefore, we define the atomic level of disease to be the cell. The reason for this is threefold. First, a cell is a self contained unit. Second, each cell has its own genetic code and a discrete and bounded number of molecules within it, at any given time. Third, current high-throughput assays allow us to measure the biomolecular contents with respect to a cell or groups of cells. In practice, a cell can be sufficiently well represented by the state of its genome, transcriptome, proteome and metabolome, and its history with respect to the four components. Formally, each of the four "-omes" can be seen as an axis of the *cell space*, where the dimensions of cell space are enumerations of genome space or the levels of transcripts, proteins and metabolites along their respective axes (note that the axes are not necessarily independent). For example, measurements from microarray data can be discretized and then projected on the dimension of transcription, whereas the proteome can also incorporate the levels of post-translationally modified molecules. While such a mapping of the cell into a vector space is one-to-one, note that the particular way of mapping is less important than the fact that it is possible. Therefore, the state of the four -omes of cell *i* define a point in the cell space at time *t* as:

$$\mathbf{c}_t^i = [\mathbf{g}_t^i \quad \mathbf{t}_t^i \quad \mathbf{p}_t^i \quad \mathbf{m}_t^i]^T,$$

where $\mathbf{g}_t^i$, $\mathbf{t}_t^i$, $\mathbf{p}_t^i$ and $\mathbf{m}_t^i$ represent the axes: genome, transcriptome, proteome and metabolome, respectively, and *T* is a transpose operator (Figure 1). We represent each of the dimensions as a vector, as opposed to a scalar, in order to allow for representations that are equally powerful but more easily interpretable than the mapping into a four dimensional space. Similarly, the current state of the organism can be summarized as a collection of the states of its cells:

$$\mathbf{c}_t = [\mathbf{c}_t^1 \quad \mathbf{c}_t^2 \quad \ldots \quad \mathbf{c}_t^{n_t}],$$

where n_t is the overall number of cells at time *t*. Clearly, querying the state of the organism by directly measuring the content of all cells would likely kill them and is currently not possible. However, this representation is still useful in systematizing the measurements taken by current technology. For example, querying the content of the human plasma corresponds to a measurement that is a function of the cells in an organism as represented by some mapping $h(\mathbf{c}_t)$. In other words, this representation can be easily generalized into a more practical basic model, for example, based on tissue instead of the cell.

We view a cell's continuous functioning as a trajectory through the cell space, $\mathbf{c}_{[0,t]}^i$, driven not only from the cell's existing traits, but also from the external cellular environment[3] $\mathbf{e}_t^i$ and its trajectory $\mathbf{e}_{[0,t]}^i$. This relationship can be generally expressed as:

$$\mathbf{c}_{t+1} = f(\mathbf{c}_{[0,t]}, \mathbf{e}_{[0,t]}),$$

where function *f* can be thought of as being dependent on the various structural and functional properties of molecules, protein-protein interactions, cell-cell communication, *etc*.

We further establish a mapping from the cell space, a space constructed from a fixed number of experimentally available cellular traits, to a *symptom space* by considering an aggregate of such cells:

$$\mathbf{s}_t = g(\mathbf{c}_t, \mathbf{e}_t),$$

where $\mathbf{s}_t = [s_t^1 \quad s_t^2 \quad \ldots \quad s_t^l]^T$ is a vector of measurable symptoms, such as bodily temperature, pain, or presence of a tumor.

Under this version of the model, the cell is a dynamic system with its space consisting of four dimensions: genomic, transcriptomic, proteomic and metabolomic. We arrived at these dimensions since they match well with the currently available technologies (*e.g.* gene expression microarrays, tandem mass spectrometry). However, the model is not limited to the current technologies and is flexible enough to handle dimensions that may arise in the future. This system is illustrated in Figure 1.

To try to formalize the notion of a biomarker, let us consider three different regions of the cell space: health-space, malfunction-space and dead-space. Health-space is a set of regions in which the cell functions normally. Malfunction-space, in contrast, contains the regions in which there is some "deviation" from a cell's programmed routine. For example, different diseases can correspond to different regions in the cell space, whereas overlapping regions correspond to situations where more than one disease is present. Dead-space, on the other hand, represents regions in which cells simply cannot exist—either through demise or not being genetically feasible. A group of cells present in their malfunction space can now be used to define *disease* on a cellular level; this disease is then mapped to a set of common symptoms. Under such a formulation, the notion of a biomarker, which is currently mostly found by statistical comparisons between two groups of samples (healthy and disease), would simply become a region in the cell space. Clearly, many genetic studies can associate probabilities of contracting a disease based on the genotype and haplotype information, which corresponds to taking only one axis of the cell space into account. Therefore, including other cellular information should, in principle, contribute to a more accurate disease biomarker and facilitate early diagnosis.

In addition to biomarker discovery and diagnosis, this approach would also be advantageous for improving prognosis and understanding of disease progression. Given large amounts of data, function f that maps a cell state into the next cell state, could be learned using computational approaches and would enable us to anticipate whether a cell (tissue) is heading towards or away from the malfunction (disease) space. Similarly, learning of function g would enable us to follow this progression at a symptom level. Note that f and g could be learned in a way to incorporate prior knowledge about a biological system.

Therefore, the major objective of disease informatics can be formulated as labeling the cell space and learning the mappings f and g. As mentioned, such tasks would incorporate anticipation of moving into the malfunction-space and undertaking preventive treatment or finding ways of moving out of a malfunction-space. Typically, physicians begin with a symptom vector $\mathbf{s}_t$ and invert the mapping to essentially conditionally guess (in the statistical sense) the cell vectors $\mathbf{c}_t$ and their relationships to disease-space. Directly learning functions f and g, however, would enable a data-driven approach in which scientists would better understand how to perturb the system in order to treat causes of disease and to enable personalized medicine.

One of the strengths of the above-mentioned model is that it takes a cell-based approach, much as the cytomics approaches do (184-187). It is inherently personalized and it moves the problem toward the experimentally measurable cellular traits away from the symptom space. We note that this model does not undermine the reductionist studies of the molecular basis of disease. In fact, such approaches will remain to be critical for diseases that are governed by one dominant factor or a small number of factors, (*e.g.* a structural defect in protein structure) and are incorporated into our model through functions f and g, or direct labeling of the malfunction space on a cellular level. In addition, this model emphasizes frequent measurements of cell states and understanding of the cellular environment and exploits the power of statistical and physical models to ultimately bring about computational models that could effectively contribute to a personalized approach to predictive medicine. While a comprehensive analysis and understanding of disease is possible, this model also suggests that there may be a long way to go until true predictive medicine is achieved.

5. SUMMARY

In this paper we reviewed a number of computational approaches used in the study of human disease. Although our coverage is broad, it is not complete. Several other disciplines that study disease fall outside the discussion of this paper, but should be mentioned to draw attention to their importance. These include, for example, genome-wide association studies, pharmacogenomics and cytomics. For a discussion on genome-wide association studies, we refer the reader to (58), for pharmacogenomics to (188) and for cytomics to (184-187).

Most of the current computational approaches that are dedicated to understanding disease provide valuable insights, but as such they may still be inadequate for a thorough and general understanding of many complex diseases. As a result, their impact on clinical studies have been limited. We suggest that while the community should continue to develop niche techniques in various subdisciplines, a more comprehensive data integration approach should be considered to more accurately learn the disease space of cells. One such approach, outlined in Section 4, will provide a more robust model of physical symptoms as they relate to a disease and provide clinicians a means to help with diagnosis and prognosis, while also providing a framework to better understand drug interactions in the cell. Although experimental data are currently limited, the amount of data in the near future will only increase, thus adding to the gap between the amount of data and our understanding of complex disease. This model has been proposed to take advantage of the computational power, experimental data and statistical and physical techniques developed over the past several decades, with an eye to the future. We hope that the proposed model provides a reasonable framework and provokes discussion of how to better design studies that take advantage of our current and future experimental data and computational methods.

6. ACKNOWLEDGEMENTS

We thank Matthew W. Hahn, Brian D. Eads and Haixu Tang of Indiana University for their comments. This work is partially supported by the NSF award DBI-0644017 to PR.

7. REFERENCES

1. Collins, F. S., E. D. Green, A. E. Guttmacher & M. S. Guyer: A vision for the future of genomics research. *Nature*, 422, 835-47 (2003)
2. Zerhouni, E.: The NIH roadmap. *Science*, 302, 63-65 (2003)
3. Dunker, A. K. & Z. Obradovic: The protein trinity - linking function and disorder. *Nat. Biotechnol.*, 19, 805-806 (2001)
4. Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin & G. Sherlock: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25, 25-29 (2000)
5. Ideker, T., T. Galitski & L. Hood: A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2, 343-72 (2001)
6. Ryan, D. P. & J. M. Matthews: Protein-protein interactions in human disease. *Curr Opin Struct Biol*, 15, 441-6 (2005)
7. Oti, M. & H. G. Brunner: The modular nature of genetic diseases. *Clin Genet*, 71, 1-11 (2007)
8. Kann, M. G.: Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform* (2007)
9. Lussier, Y. A. & Y. Liu: Computational approaches to phenotyping: high-throughput phenomics. *Proc Am Thorac Soc*, 4, 18-25 (2007)
10. Loscalzo, J., I. Kohane & A. L. Barabasi: Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol*, 3, 124 (2007)
11. Knibbs, G. H.: The international classification of disease and causes of death and its revision. *Medical journal of Australia*, 1, 2-12 (1929)
12. International statistical classification of diseases and health related problems: ICD-10. World Health Organization, Geneva, Switzerland (2005)
13. Stearns, M. Q., C. Price, K. A. Spackman & A. Y. Wang: SNOMED clinical terms: overview of the development process and project status. *Proc AMIA Symp*662-6 (2001)
14. Lindberg, D. A., B. L. Humphreys & A. T. McCray: The Unified Medical Language System. *Methods Inf Med*, 32, 281-91 (1993)
15. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32, D267-70 (2004)
16. Lussier, Y., T. Borlawsky, D. Rappaport, Y. Liu & C. Friedman: PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput*64-75 (2006)
17. Friedman, C., L. Shagina, Y. Lussier & G. Hripcsak: Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*, 11, 392-402 (2004)
18. Lussier, Y. A. & J. Li: Terminological mapping for high throughput comparative biology of phenotypes. *Pac Symp Biocomput*202-13 (2004)
19. Bedell, M. A., N. A. Jenkins & N. G. Copeland: Mouse models of human disease. Part I: techniques and resources for genetic analysis in mice. *Genes Dev*, 11, 1-10 (1997)
20. Bedell, M. A., D. A. Largaespada, N. A. Jenkins & N. G. Copeland: Mouse models of human disease. Part II: recent progress and future directions. *Genes Dev*, 11, 11-43 (1997)
21. Bier, E.: Drosophila, the golden bug, emerges as a tool for human genetics. *Nat Rev Genet*, 6, 9-23 (2005)
22. Culetto, E. & D. B. Sattelle: A role for Caenorhabditis elegans in understanding the function and interactions of human disease genes. *Hum Mol Genet*, 9, 869-77 (2000)
23. Remm, M., C. E. Storm & E. L. Sonnhammer: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314, 1041-52 (2001)
24. O'Brien, K. P., I. Westerlund & E. L. Sonnhammer: OrthoDisease: a database of human disease orthologs. *Hum Mutat*, 24, 112-9 (2004)
25. Mushegian, A. R., D. E. Bassett, Jr., M. S. Boguski, P. Bork & E. V. Koonin: Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc Natl Acad Sci U S A*, 94, 5831-6 (1997)
26. Lopez-Bigas, N. & C. A. Ouzounis: Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*, 32, 3108-14 (2004)
27. Uetz, P., L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields & J. M. Rothberg: A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae. Nature*, 403, 623-627 (2000)
28. Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant & J. M. Rothberg: A protein interaction map of Drosophila melanogaster. *Science*, 302, 1727-36 (2003)
29. Chintapalli, V. R., J. Wang & J. A. Dow: Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nat Genet*, 39, 715-20 (2007)
30. Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker & J. B. Hogenesch: A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101, 6062-7 (2004)

31. Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori & Y. Sakaki: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.*, 98, 4569-74. (2001)
32. Yu, H., N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J. D. Han, N. Bertin, S. Chung, M. Vidal & M. Gerstein: Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*, 14, 1107-18 (2004)
33. Matthews, L. R., P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent & M. Vidal: Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*, 11, 2120-6 (2001)
34. Walhout, A. J., R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg & M. Vidal: Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, 287, 116-22 (2000)
35. Brown, K. R. & I. Jurisica: Online Predicted Human Interaction Database. In, (2005)
36. Brown, K. R. & I. Jurisica: Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol*, 8, R95 (2007)
37. Gibbs, R. A., J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock, E. R. Mardis, K. A. Remington, R. L. Strausberg, J. 38. C. Venter, R. K. Wilson, M. A. Batzer, C. D. Bustamante, E. E. Eichler, M. W. Hahn, R. C. Hardison, K. D. Makova, W. Miller, A. Milosavljevic, R. E. Palermo, A. Siepel, J. M. Sikela, T. Attaway, S. Bell, K. E. Bernard, C. J. Buhay, M. N. Chandrabose, M. Dao, C. Davis, K. D. Delehaunty, Y. Ding, H. H. Dinh, S. Dugan-Rocha, L. A. Fulton, R. A. Gabisi, T. T. Garner, J. Godfrey, A. C. Hawes, J. Hernandez, S. Hines, M. Holder, J. Hume, S. N. Jhangiani, V. Joshi, Z. M. Khan, E. F. Kirkness, A. Cree, R. G. Fowler, S. Lee, L. R. Lewis, Z. Li, Y. S. Liu, S. M. Moore, D. Muzny, L. V. Nazareth, D. N. Ngo, G. O. Okwuonu, G. Pai, D. Parker, H. A. Paul, C. Pfannkoch, C. S. Pohl, Y. H. Rogers, S. J. Ruiz, A. Sabo, J. Santibanez, B. W. Schneider, S. M. Smith, E. Sodergren, A. F. Svatek, T. R. Utterback, S. Vattathil, W. Warren, C. S. White, A. T. Chinwalla, Y. Feng, A. L. Halpern, L. W. Hillier, X. Huang, P. Minx, J. O. Nelson, K. H. Pepin, X. Qin, G. G. Sutton, E. Venter, B. P. Walenz, J. W. Wallis, K. C. Worley, S. P. Yang, S. M. Jones, M. A. Marra, M. Rocchi, J. E. Schein, R. Baertsch, L. Clarke, M. Csuros, J. Glasscock, R. A. Harris, P. Havlak, A. R. Jackson, H. Jiang, Y. Liu, D. N. Messina, Y. Shen, H. X. Song, T. Wylie, L. Zhang, E. Birney, K. Han, M. K. Konkel, J. Lee, A. F. Smit, B. Ullmer, H. Wang, J. Xing, R. Burhans, Z. Cheng, J. E. Karro, J. Ma, B. Raney, X. She, M. J. Cox, J. P. Demuth, L. J. Dumas, S. G. Han, J. Hopkins, A. Karimpour-Fard, Y. H. Kim, J. R. Pollack, T. Vinar, C. Addo-Quaye, J. Degenhardt, A. Denby, M. J. Hubisz, A. Indap, C. Kosiol, B. T. Lahn, H. A. Lawson, A. Marklein, R. Nielsen, E. J. Vallender, A. G. Clark, B. Ferguson, R. D. Hernandez, K. Hirani, H. Kehrer-Sawatzki, J. Kolb, S. Patil, L. L. Pu, Y. Ren, D. G. Smith, D. A. Wheeler, I. Schenck, E. V. Ball, R. Chen, D. N. Cooper, B. Giardine, F. Hsu, W. J. Kent, A. Lesk, D. L. Nelson, E. O'Brien W, K. Prufer, P. D. Stenson, J. C. Wallace, H. Ke, X. M. Liu, P. Wang, A. P. Xiang, F. Yang, G. P. Barber, D. Haussler, D. Karolchik, A. D. Kern, R. M. Kuhn, K. E. Smith & A. S. Zwieg: Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316, 222-34 (2007)
38. Ostedgaard, L. S., C. S. Rogers, Q. Dong, C. O. Randak, D. W. Vermeer, T. Rokhlina, P. H. Karp & M. J. Welsh: Processing and function of CFTR-DeltaF508 are species-dependent. *Proc Natl Acad Sci U S A*, 104, 15370-5 (2007)
39. Ross, D. T., U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein & P. O. Brown: Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, 24, 227-35 (2000)
40. Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M.

Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh & X. Zhu: The sequence of the human genome. *Science*, 291, 1304-51. (2001)

41. Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi & Y. J. Chen: Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921 (2001)

42. Jimenez-Sanchez, G., B. Childs & D. Valle: Human disease genes. *Nature*, 409, 853-5 (2001)

43. Karlin, S., L. Brocchieri, A. Bergman, J. Mrazek & A. J. Gentles: Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci U S A*, 99, 333-8 (2002)

44. Smith, N. G. & A. Eyre-Walker: Human disease genes: patterns and predictions. *Gene*, 318, 169-75 (2003)

45. Bortoluzzi, S., C. Romualdi, A. Bisognin & G. A. Danieli: Disease genes and intracellular protein networks. *Physiol Genomics*, 15, 223-7 (2003)

46. Huang, H., E. E. Winter, H. Wang, K. G. Weinstock, H. Xing, L. Goodstadt, P. D. Stenson, D. N. Cooper, D. Smith, M. M. Alba, C. P. Ponting & K. Fechtel: Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol*, 5, R47 (2004)

47. Tu, Z., L. Wang, M. Xu, X. Zhou, T. Chen & F. Sun: Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, 7, 31 (2006)

48. Goh, K. I., M. E. Cusick, D. Valle, B. Childs, M. Vidal & A. L. Barabasi: The human disease network. *Proc Natl Acad Sci U S A*, 104, 8685-90 (2007)

49. Cheng, Y., T. LeGall, C. J. Oldfield, A. K. Dunker & V. N. Uversky: Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry*, 45, 10448-60 (2006)

50. Iakoucheva, L. M., C. J. Brown, J. D. Lawson, Z. Obradovic & A. K. Dunker: Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, 323, 573-584 (2002)

51. Thomas, M. A., B. Weston, M. Joseph, W. Wu, A. Nekrutenko & P. J. Tonellato: Evolutionary dynamics of oncogenes and tumor suppressor genes: higher intensities of purifying selection than other genes. *Mol Biol Evol*, 20, 964-8 (2003)

52. Jonsson, P. F. & P. A. Bates: Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22, 2291-7 (2006)

53. Furney, S. J., D. G. Higgins, C. A. Ouzounis & N. Lopez-Bigas: Structural and functional properties of genes involved in human cancer. *BMC Genomics*, 7, 3 (2006)

54. Adie, E. A., R. R. Adams, K. L. Evans, D. J. Porteous & B. S. Pickard: Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6, 55 (2005)

55. Xu, J. & Y. Li: Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 22, 2800-5 (2006)

56. Risch, N. J.: Searching for genetic determinants in the new millennium. *Nature*, 405, 847-56 (2000)

57. Glazier, A. M., J. H. Nadeau & T. J. Aitman: Finding genes that underlie complex traits. *Science*, 298, 2345-9 (2002)

58. Hirschhorn, J. N. & M. J. Daly: Genome-wide

association studies for common diseases and complex traits. *Nat Rev Genet*, 6, 95-108 (2005)
59. Chen, J. Y., C. Shen & A. Y. Sivachenko: Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac Symp Biocomput*, 11, 367-378 (2006)
60. Oti, M., B. Snel, M. A. Huynen & H. G. Brunner: Predicting disease genes using protein-protein interactions. *J Med Genet*, 43, 691-8 (2006)
61. Gonzalez, G., J. C. Uribe, L. Tari, C. Brophy & C. Baral: Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity. *Pac Symp Biocomput*, 12, 28-39 (2007)
62. Lage, K., E. O. Karlberg, Z. M. Storling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, Y. Moreau & S. Brunak: A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25, 309-16 (2007)
63. Perez-Iratxeta, C., P. Bork & M. A. Andrade: Association of genes to genetically inherited diseases using data mining. *Nat Genet*, 31, 316-9 (2002)
64. Perez-Iratxeta, C., M. Wjst, P. Bork & M. A. Andrade: G2D: a tool for mining genes associated with disease. *BMC Genet*, 6, 45 (2005)
65. Seki, K. & J. Mostafa: Discovering implicit associations between genes and hereditary diseases. *Pac Symp Biocomput*, 12, 316-327 (2007)
66. Hristovski, D., B. Peterlin, J. A. Mitchell & S. M. Humphrey: Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*, 74, 289-98 (2005)
67. Turner, F. S., D. R. Clutterbuck & C. A. Semple: POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*, 4, R75 (2003)
68. Adie, E. A., R. R. Adams, K. L. Evans, D. J. Porteous & B. S. Pickard: SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22, 773-4 (2006)
69. George, R. A., J. Y. Liu, L. L. Feng, R. J. Bryson-Richardson, D. Fatkin & M. A. Wouters: Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* (2006)
70. Freudenberg, J. & P. Propping: A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18 Suppl 2, S110-5 (2002)
71. Rossi, S., D. Masotti, C. Nardini, E. Bonora, G. Romeo, E. Macii, L. Benini & S. Volinia: TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res*, 34, W285-92 (2006)
72. Franke, L., H. Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen & C. Wijmenga: Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78, 1011-25 (2006)
73. Aerts, S., D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L. C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet & Y. Moreau: Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24, 537-44 (2006)
74. De Bie, T., L. C. Tranchevent, L. M. van Oeffelen & Y. Moreau: Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23, i125-32 (2007)
75. Costello, J. C., J. E. Buchanan-Carter, M. M. Dalkilic & J. Andrews: Using *Drosophila melanogaster* data to discover disease-related protein interactions in humans. *Proc IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 469-475 (2007).
76. Korbel, J. O., T. Doerks, L. J. Jensen, C. Perez-Iratxeta, S. Kaczanowski, S. D. Hooper, M. A. Andrade & P. Bork: Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol*, 3, e134 (2005)
77. Butte, A. J. & I. S. Kohane: Creation and implications of a phenome-genome network. *Nat Biotechnol*, 24, 55-62 (2006)
78. Braun, T. A., S. P. Shankar, S. Davis, B. O'Leary, T. E. Scheetz, A. F. Clark, V. C. Sheffield, T. L. Casavant & E. M. Stone: Prioritizing regions of candidate genes for efficient mutation screening. *Hum Mutat*, 27, 195-200 (2006)
79. Hart, G. T., A. K. Ramani & E. M. Marcotte: How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7, 120 (2006)
80. Couzin, J.: Microarray data reproduced, but some concerns remain. *Science*, 313, 1559 (2006)
81. Frayling, T. M., N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. Perry, K. S. Elliott, H. Lango, N. W. Rayner, B. Shields, L. W. Harries, J. C. Barrett, S. Ellard, C. J. Groves, B. Knight, A. M. Patch, A. R. Ness, S. Ebrahim, D. A. Lawlor, S. M. Ring, Y. Ben-Shlomo, M. R. Jarvelin, U. Sovio, A. J. Bennett, D. Melzer, L. Ferrucci, R. J. Loos, I. Barroso, N. J. Wareham, F. Karpe, K. R. Owen, L. R. Cardon, M. Walker, G. A. Hitman, C. N. Palmer, A. S. Doney, A. D. Morris, G. D. Smith, A. T. Hattersley & M. I. McCarthy: A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316, 889-94 (2007)
82. Mooney, S. D.: Bioinformatics approaches and resources for signle nucleotide polymorphism functional analysis. *Brief Bioinform*, 6, 44-56 (2005)
83. Peltonen, L. & V. A. McKusick: Genomics and medicine. Dissecting human disease in the postgenomic era. *Science*, 291, 1224-9 (2001)
84. Hamosh, A., A. F. Scott, J. Amberger, D. Valle & V. A. McKusick: Online Mendelian Inheritance in Man (OMIM). *Hum Mutat*, 15, 57-61 (2000)
85. Sunyaev, S., V. Ramensky & P. Bork: Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*, 16, 198-200 (2000)
86. Sunyaev, S., V. Ramensky, I. Koch, W. Lathe, 3rd, A. S. Kondrashov & P. Bork: Prediction of deleterious human alleles. *Hum Mol Genet*, 10, 591-7 (2001)
87. Saunders, C. T. & D. Baker: Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol*, 322, 891-901 (2002)
88. Ng, P. C. & S. Henikoff: SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31, 3812-4 (2003)
89. Karchin, R., M. Diekhans, L. Kelly, D. J. Thomas, U. Pieper, N. Eswar, D. Haussler & A. Sali: LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, 21, 2814-20

(2005)
90. Ferrer-Costa, C., M. Orozco & X. de la Cruz: Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol*, 315, 771-86 (2002)
91. Stitziel, N. O., Y. Y. Tseng, D. Pervouchine, D. Goddeau, S. Kasif & J. Liang: Structural location of disease-associated single-nucleotide polymorphisms. *J Mol Biol*, 327, 1021-30 (2003)
92. Yue, P., Z. Li & J. Moult: Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol*, 353, 459-73 (2005)
93. Ye, Y., Z. Li & A. Godzik: Modeling and analyzing three-dimensional structures of human disease proteins. *Pac Symp Biocomput*, 11, 439-450 (2006)
94. Yue, P. & J. Moult: Identification and analysis of deleterious human SNPs. *J Mol Biol*, 356, 1263-74 (2006)
95. Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski & K. Sirotkin: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29, 308-11 (2001)
96. Ng, P. C. & S. Henikoff: Accounting for human polymorphisms predicted to affect protein function. *Genome Res*, 12, 436-46 (2002)
97. Chasman, D. & R. M. Adams: Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol*, 307, 683-706 (2001)
98. Wang, Z. & J. Moult: SNPs, protein structure, and disease. *Hum Mutat*, 17, 263-70 (2001)
99. Schon, O., A. Friedler, M. Bycroft, S. M. Freund & A. R. Fersht: Molecular mechanism of the interaction between MDM2 and p53. *J Mol Biol*, 323, 491-501 (2002)
100. Fuxreiter, M., I. Simon, P. Friedrich & P. Tompa: Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol*, 338, 1015-26 (2004)
101. Mohan, A., C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker & V. N. Uversky: Analysis of molecular recognition features (MoRFs). *J Mol Biol*, 362, 1043-59 (2006)
102. Vacic, V., C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky & A. K. Dunker: Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res*, 6, 2351-66 (2007)
103. Garner, E., P. Romero, A. K. Dunker, C. Brown & Z. Obradovic: Predicting binding regions within disordered proteins. *Genome Inform. Ser. Workshop Genome Inform.*, 10, 41-50 (1999)
104. Oldfield, C. J., Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky & A. K. Dunker: Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry*, 44, 12454-70 (2005)
105. Radivojac, P., S. Vucetic, T. R. O'Connor, V. N. Uversky, Z. Obradovic & A. K. Dunker: Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins*, 63, 398-410 (2006)
106. Uversky, V. N. & A. L. Fink: Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim Biophys Acta*, 1698, 131-53 (2004)
107. Linding, R., J. Schymkowitz, F. Rousseau, F. Diella & L. Serrano: A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J Mol Biol*, 342, 345-53 (2004)
108. Fernandez, A. & R. Scott: Dehydron: a structurally encoded signal for protein interaction. *Biophys J*, 85, 1914-28 (2003)
109. Fernandez, A.: Functionality of wrapping defects in soluble proteins: what cannot be kept dry must be conserved. *J Mol Biol*, 337, 477-83 (2004)
110. Fernandez, A. & R. S. Berry: Molecular dimension explored in evolution to promote proteomic complexity. *Proc Natl Acad Sci U S A*, 101, 13460-5 (2004)
111. Fernandez, A., J. Kardos, L. R. Scott, Y. Goto & R. S. Berry: Structural defects and the diagnosis of amyloidogenic propensity. *Proc Natl Acad Sci U S A*, 100, 6446-51 (2003)
112. Fisher, M. T.: Proline to the rescue. *Proc Natl Acad Sci U S A*, 103, 13265-6 (2006)
113. Dobson, C. M.: The structural basis of protein folding and its links with human disease. *Philos Trans R Soc Lond B Biol Sci*, 356, 133-45 (2001)
114. Dobson, C. M. & M. Karplus: The fundamentals of protein folding: bringing together theory and experiment. *Curr Opin Struct Biol*, 9, 92-101 (1999)
115. Karplus, M. & J. A. McCammon: Molecular dynamics simulations of biomolecules. *Nat Struct Biol*, 9, 646-52 (2002)
116. Lucent, D., V. Vishal & V. S. Pande: Protein folding under confinement: a role for solvent. *Proc Natl Acad Sci U S A*, 104, 10430-4 (2007)
117. Huang, Q., C. L. Chen & A. Herrmann: Bilayer conformation of fusion peptide of influenza virus hemagglutinin: a molecular dynamics simulation study. *Biophys J*, 87, 14-22 (2004)
118. Lague, P., B. Roux & R. W. Pastor: Molecular dynamics simulations of the influenza hemagglutinin fusion peptide in micelles and bilayers: conformational analysis of peptide and lipids. *J Mol Biol*, 354, 1129-41 (2005)
119. Kasson, P. M. & V. S. Pande: Predicting structure and dynamics of loosely-ordered protein complexes: influenza hemagglutinin fusion peptide. *Pac Symp Biocomput*, 12, 40-50 (2007)
120. Shamsir, M. S. & A. R. Dalby: One gene, two diseases and three conformations: molecular dynamics simulations of mutants of human prion protein at room temperature and elevated temperatures. *Proteins*, 59, 275-90 (2005)
121. Raffa, D. F. & A. Rauk: Molecular dynamics study of the beta amyloid peptide of Alzheimer's disease and its divalent copper complexes. *J Phys Chem B*, 111, 3789-99 (2007)
122. Schena, M., D. Shalon, R. W. Davis & P. O. Brown: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270, 467-70 (1995)
123. Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield & E. S. Lander: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-7 (1999)
124. Rhodes, D. R., T. R. Barrette, M. A. Rubin, D. Ghosh

& A. M. Chinnaiyan: Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*, 62, 4427-33 (2002)
125. Matsuzaki, H., S. Dong, H. Loi, X. Di, G. Liu, E. Hubbell, J. Law, T. Berntsen, M. Chadha, H. Hui, G. Yang, G. C. Kennedy, T. A. Webster, S. Cawley, P. S. Walsh, K. W. Jones, S. P. Fodor & R. Mei: Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods*, 1, 109-11 (2004)
126. Matsuzaki, H., H. Loi, S. Dong, Y. Y. Tsai, J. Fang, J. Law, X. Di, W. M. Liu, G. Yang, G. Liu, J. Huang, G. C. Kennedy, T. B. Ryder, G. A. Marcus, P. S. Walsh, M. D. Shriver, J. M. Puck, K. W. Jones & R. Mei: Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res*, 14, 414-25 (2004)
127. Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray & D. G. Albertson: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, 20, 207-11 (1998)
128. Pollack, J. R., T. Sorlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Borresen-Dale & P. O. Brown: Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99, 12963-8 (2002)
129. Miklos, G. L. & R. Maleszka: Microarray reality checks in the context of a complex disease. *Nat Biotechnol*, 22, 615-21 (2004)
130. Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown & L. M. Staudt: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403, 503-11 (2000)
131. van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards & S. H. Friend: Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415, 530-6 (2002)
132. Furey, T. S., N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer & D. Haussler: Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16, 906-14 (2000)
133. Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler & L. C. Groop: PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34, 267-73 (2003)
134. Hosack, D. A., G. Dennis, Jr., B. T. Sherman, H. C. Lane & R. A. Lempicki: Identifying biological themes within lists of genes with EASE. *Genome Biol*, 4, R70 (2003)
135. Lee, H. K., W. Braynen, K. Keshav & P. Pavlidis: ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, 6, 269 (2005)
136. Dhanasekaran, S. M., T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin & A. M. Chinnaiyan: Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412, 822-6 (2001)
137. Magee, J. A., T. Araki, S. Patil, T. Ehrig, L. True, P. A. Humphrey, W. J. Catalona, M. A. Watson & J. Milbrandt: Expression profiling reveals hepsin overexpression in prostate cancer. *Cancer Res*, 61, 5692-6 (2001)
138. Welsh, J. B., L. M. Sapinoso, A. I. Su, S. G. Kern, J. Wang-Rodriguez, C. A. Moskaluk, H. F. Frierson, Jr. & G. M. Hampton: Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res*, 61, 5974-8 (2001)
139. Luo, J., D. J. Duggan, Y. Chen, J. Sauvageot, C. M. Ewing, M. L. Bittner, J. M. Trent & W. B. Isaacs: Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res*, 61, 4683-8 (2001)
140. Rhodes, D. R., S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B. B. Briggs, T. R. Barrette, M. J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh & A. M. Chinnaiyan: Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, 9, 166-80 (2007)
141. Baranova, A., K. Schlauch, S. Gowder, R. Collantes, V. Chandhoke & Z. M. Younossi: Microarray technology in the study of obesity and non-alcoholic fatty liver disease. *Liver Int*, 25, 1091-6 (2005)
142. Eaves, I. A., L. S. Wicker, G. Ghandour, P. A. Lyons, L. B. Peterson, J. A. Todd & R. J. Glynne: Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. *Genome Res*, 12, 232-43 (2002)
143. Revel, A. T., A. M. Talaat & M. V. Norgard: DNA microarray analysis of differential gene expression in Borrelia burgdorferi, the Lyme disease spirochete. *Proc Natl Acad Sci U S A*, 99, 1562-7 (2002)
144. Albertson, D. G. & D. Pinkel: Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet*, 12 Spec No 2, R145-52 (2003)
145. Vissers, L. E., C. M. van Ravenswaaij, R. Admiraal, J. A. Hurst, B. B. de Vries, I. M. Janssen, W. A. van der Vliet, E. H. Huys, P. J. de Jong, B. C. Hamel, E. F. Schoenmakers, H. G. Brunner, J. A. Veltman & A. G. van Kessel: Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat Genet*, 36, 955-7 (2004)
146. Vissers, L. E., J. A. Veltman, A. G. van Kessel & H. G. Brunner: Identification of disease genes by whole genome CGH arrays. *Hum Mol Genet*, 14 Spec No. 2, R215-23 (2005)
147. Teh, M. T., D. Blaydon, T. Chaplin, N. J. Foot, S. Skoulakis, M. Raghavan, C. A. Harwood, C. M. Proby, M. P. Philpott, B. D. Young & D. P. Kelsell: Genomewide single nucleotide polymorphism microarray mapping in basal cell carcinomas unveils uniparental disomy as a key

somatic event. *Cancer Res*, 65, 8597-603 (2005)
148. LaFramboise, T., B. A. Weir, X. Zhao, R. Beroukhim, C. Li, D. Harrington, W. R. Sellers & M. Meyerson: Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput Biol*, 1, e65 (2005)
149. Bandyopadhyay, S., R. Kelley & T. Ideker: Discovering regulated networks during HIV-1 latency and reactivation. *Pac Symp Biocomput*, 11, 354-366 (2006)
150. Valdivia-Granda, W. A., M. G. Kann & J. Malaga: Transcriptional interactions during smallpox infection and identification of early infection biomarkers. *Pac Symp Biocomput*, 12, 100-111 (2007)
151. Aebersold, R. & M. Mann: Mass spectrometry-based proteomics. *Nature*, 422, 198-207 (2003)
152. Yates, J. R., 3rd: Mass spectral analysis in proteomics. *Annu Rev Biophys Biomol Struct*, 33, 297-316 (2004)
153. Russell, S. A., W. Old, K. A. Resing & L. Hunter: Proteomic informatics. *Int Rev Neurobiol*, 61, 127-57 (2004)
154. Hanash, S.: Disease proteomics. *Nature*, 422, 226-32 (2003)
155. Diamandis, E. P.: Is early detection of cancer with serum biomarkers or proteomics profiling feasible? *AACR Education Book*, 2007, 129-132 (2007)
156. Petricoin, E. F., A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn & L. A. Liotta: Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359, 572-7 (2002)
157. Diamandis, E. P.: Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. *Mol Cell Proteomics*, 3, 367-378 (2004)
158. Wu, B., T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams & H. Zhao: Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19, 1636-43 (2003)
159. Perkins, D. N., D. J. Pappin, D. M. Creasy & J. S. Cottrell: Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20, 3551-67 (1999)
160. Yates, J. R., 3rd, J. K. Eng, A. L. McCormack & D. Schieltz: Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, 67, 1426-36 (1995)
161. Geer, L. Y., S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi & S. H. Bryant: Open mass spectrometry search algorithm. *J Proteome Res*, 3, 958-64 (2004)
162. Craig, R. & R. C. Beavis: TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20, 1466-7 (2004)
163. Nesvizhskii, A. I. & R. Aebersold: Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, 4, 1419-1440 (2005)
164. Nesvizhskii, A. I., A. Keller, E. Kolker & R. Aebersold: A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75, 4646-58 (2003)
165. Alves, P., R. J. Arnold, M. V. Novotny, P. Radivojac, J. P. Reilly & H. Tang: Advancements in protein identification from shotgun proteomics using predicted peptide detectability. *Pac Symp Biocomput*, 12, 409-420 (2007)
166. Zhang, B., M. C. Chambers & D. L. Tabb: Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res*, 6, 3549-57 (2007)
167. Higgs, R. E., M. D. Knierman, V. Gelfanova, J. P. Butler & J. E. Hale: Comprehensive label-free method for the relative quantification of proteins from biological samples. *J Proteome Res*, 4, 1442-50 (2005)
168. Tang, H., R. J. Arnold, P. Alves, Z. Xun, D. E. Clemmer, M. V. Novotny, J. P. Reilly & P. Radivojac: A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics*, 22, e481-e488 (2006)
169. Lu, P., C. Vogel, R. Wang, X. Yao & E. M. Marcotte: Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25, 117-124 (2007)
170. Mann, M. & O. N. Jensen: Proteomic analysis of post-translational modifications. *Nat Biotechnol*, 21, 255-61 (2003)
171. Mann, M., S. E. Ong, M. Gronborg, H. Steen, O. N. Jensen & A. Pandey: Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol.*, 20, 261-268 (2002)
172. Tang, H., Y. Mechref & M. V. Novotny: Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics*, 21 Suppl 1, i431-9 (2005)
173. Wishart, D. S., D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. Macinnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel & L. Querengesser: HMDB: the Human Metabolome Database. *Nucleic Acids Res*, 35, D521-6 (2007)
174. Kenny, L. C., W. B. Dunn, D. I. Ellis, J. Myers, P. N. Baker, GOPEC Conortium & D. B. Kell: Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. *Metabolomics*, 1, 227-234 (2005)
175. Schnackenberg, L., R. D. Beger & Y. Dragan: NMR-based metabonomic evaluation of livers from rats chronically treated with tamoxifen, mestranol, and phenobarbital. *Metabolomics*, 1, 87-94 (2005)
176. Oostendorp, M., U. F. Engelke, M. A. Willemsen & R. A. Wevers: Diagnosing inborn errors of lipid metabolism with proton nuclear magnetic resonance spectroscopy. *Clin Chem*, 52, 1395-405 (2006)
177. Sysi-Aho, M., A. Vehtari, V. R. Velagapudi, J. Westerbacka, L. Yetukuri, R. Bergholm, M. R. Taskinen, H. Yki-Jarvinen & 178. M. Oresic: Exploring the lipoprotein composition using Bayesian regression on serum lipidomic profiles. *Bioinformatics*, 23, i519-28 (2007)
179. Zhang, X., D. Wei, Y. Yap, L. Li, S. Guo & F. Chen: Mass spectrometry-based "omics" technologies in cancer diagnostics. *Mass Spectrom Rev*, 26, 403-31 (2007)
180. Wishart, D. S.: Current Progress in computational metabolomics. *Brief Bioinform* (2007)

181. Hertz, H. S., R. A. Hites & K. Biemann: Identification of mass spectra by computer searching a file of known spectra. *Anal. Chem.*, 43, 681-691 (1971)
182. Pesyna, G. M., R. Venkataraghavan, H. E. Dayrnger & F. W. McLafferty: Probability based matching system using a large collection of reference mass spectra. *Anal Chem*, 48, 1362-1368 (1976)
183. Robertson, D. G., M. D. Reily & J. D. Baker: Metabonomics in pharmaceutical discovery and development. *J. Proteome Res.*, 6, 526-539 (2007)
184. Pearson, H.: Meet the human metabolome. *Nature*, 446, 8 (2007)
185. Valet, G.: Predictive medicine by cytomics: potential and challenges. *J Biol Regul Homeost Agents*, 16, 164-7 (2002)
186. Valet, G. K. & A. Tarnok: Cytomics in predictive medicine. *Cytometry B Clin Cytom*, 53, 1-3 (2003)
187. Valet, G., J. F. Leary & A. Tarnok: Cytomics--new technologies: towards a human cytome project. *Cytometry A*, 59, 167-71 (2004)
188. Kriete, A.: Cytomics in the realm of systems biology. *Cytometry A*, 68, 19-20 (2005)
189. Klein, T. E., J. T. Chang, M. K. Cho, K. L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D. E. Oliver, D. L. Rubin, F. Shafa, J. M. Stuart & R. B. Altman: Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J*, 1, 167-70 (2001)

Footnotes: [1]http://diseaseontology.sourceforge.net/, [2] One can consider drug exposure, for instance, or presence of microbial populations within an organism as an environmental factor, [3] http://www.progenetix.de/~pgscripts/progenetix

Key Words: Disease, Informatics, Gene Associations, Protein Associations, Protein Structure, Review

Send correspondence to: Dr Predrag Radivojac, School of Informatics, Indiana University, Bloomington, IN 47408, Tel: 812-856-1851, Fax: 812-856-1995, E-mail: predrag@indiana.edu

http://www.bioscience.org/current/vol13.htm