# Trends and Developments in Bioinformatics in 2010: Prospects and Perspectives

C. F. Aliferis, A. V. Alekseyenko, Y. Aphinyanaphongs, S. Brown, D. Fenyo, L. Fu, S. Shen, A. Statnikov, J. Wang
Center for Health Informatics and Bioinformatics, New York University

## Summary

***Objectives***: To survey major developments and trends in the field of Bioinformatics in 2010 and their relationships to those of previous years, with emphasis on long-term trends, on best practices, on quality of the science of informatics, and on quality of science as a function of informatics.

***Methods***: A critical review of articles in the literature of Bioinformatics over the past year.

***Results***: Our main results suggest that Bioinformatics continues to be a major catalyst for progress in Biology and Translational Medicine, as a consequence of new assaying technologies, most predominantly Next Generation Sequencing, which are changing the landscape of modern biological and medical research. These assays critically depend on bioinformatics and have led to quick growth of corresponding informatics methods development. Clinical-grade molecular signatures are proliferating at a rapid rate. However, a highly publicized incident at a prominent university showed that deficiencies in informatics methods can lead to catastrophic consequences for important scientific projects. Developing evidence-driven protocols and best practices is greatly needed given how serious are the implications for the quality of translational and basic science.

***Conclusions***: Several exciting new methods have appeared over the past 18 months, that open new roads for progress in bioinformatics methods and their impact in biomedicine. At the same time, the range of open problems of great significance is extensive, ensuring the vitality of the field for many years to come.

## Keywords

Bioinformatics, translational medicine, molecular profiles, high throughput assays, next generation sequencing

Yearb Med Inform 2011: 146-55

## Introduction

Any attempt to survey and summarize a field as diverse and large as Bioinformatics is very hard because of its volume, complexity and incredibly dynamic nature, which makes short and long term predictions risky. We therefore had to find a realistic method to identify, summarize and comment on this extraordinary body of literature in a manner that does not ignore the most essential developments and most important of all, can lead to new insights for the reader.

Our chosen methodology consists of three elements: (a) We interviewed, and invited to be co-authors of the present report, all practicing bioinformaticians at NYU Medical center. These qualified faculty members collectively support the advanced bioinformatics needs of all the NYULMC faculty (approximately 100 projects a year), support the operations of all high-throughput instruments in the Medical Center (>75 projects a year), teach Bioinformatics at the undergraduate and graduate level, lead 6 informatics method development labs, publish actively both new methods as well as methods evaluations and applications in numerous projects, and finally, are active members of all major related professional societies and participants in several highly effective national consortia and working groups covering many aspects of bioinformatics. (b) We conducted a bibliographic analysis of the field and compared the results to those reported by Kohane [1] for 2008. (c) We reviewed our consulting files and related best practices and benchmarks from our NYU best practices comprehensive consulting service (BPIC) that since 2009 supports approximately 100 frontline scientific projects every year to identify areas of continuing and emerging importance. Our goal was to ground the editorial to not only broader theoretical issues the field faces and that can be gleaned by the literature, but also to look into Bioinformatics advances from the level of real-life collaborative science that critically depends on, or is driven by, informatics advances.

In the present paper we conceptualize the modern (post Human Genome Project) field of Bioinformatics as consisting of three interrelated areas: the first area studies techniques that deal with high throughput (HT) molecular assays and produces related data (we call this "*Informatics for executing complex molecular assays*"). The second area studies methods that link molecular information to disease phenotypes, emphasizing the problems of diagnosis and treatment (this is a strongly translationally - oriented "*Informatics for knowledge discovery related to managing clinical phenotypes*"). The third area deals with discovery of knowledge about biological mechanisms (this is a more basic science - oriented "*Informatics for biological knowledge discovery*"). See Figure 1.

In general we are concerned with major trends and foundational issues. Our emphasis in the present review is more on the long-term versus the short-term trends, on best practices, on quality of the science of informatics, and on quality of science as a function of informatics.

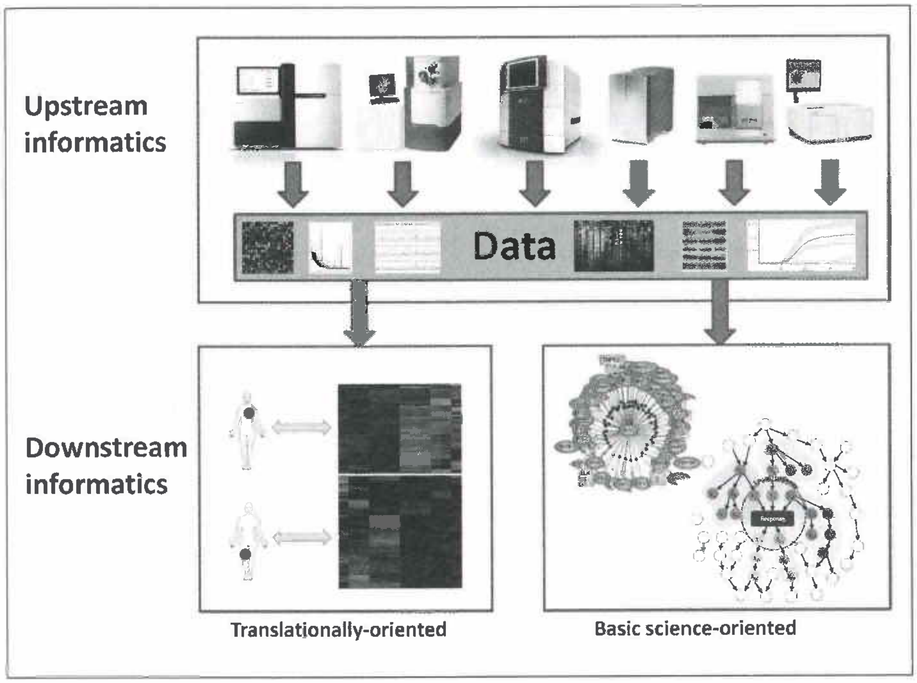

**Fig. 1** Conceptualization of Bioinformatics field used in present review

Our report is, by its very nature incomplete and limited. Omission of important papers is to be expected although we made every effort to not miss any major trend and development in broad terms. We do cite some of our work in the stated contexts without claiming or implying that any of the cited work (from our group or others) is necessarily superior to other work not cited here. Yet the cited works were instrumental in helping us first form and then articulate our perspective on what we believe are general trends and important problems to be solved. We also intended that the cited works are clear examples of the trends they were chosen to represent.

## Bibliographic Analysis of the Bioinformatics Field in 2010

In 2008, Isaac Kohane collected 10,000 bioinformatics papers and reported statistics and trends in the field [1]. We replicated this analysis in 2010. We searched the MEDLINE database with the search query "("computational biology" [MeSH Terms] OR ("computational"[All Fields] AND "biology"[All Fields]) OR "computational biology"[All Fields] OR "bioinformatics"[All Fields]) AND ("2010/01/01"[PDAT]: "2010/12/31"[PDAT])".

Comparisons between 2008 and 2010 reveal several interesting findings. In 2010, 1,630 journals published 10,991 papers by 48,210 authors, whereas In 2008, 1,478 journals were reported to have published 10,169 bioinformatics papers by 39,003 authors. Comparing these data from 2008 and 2010 shows that 10% more journals published 8% more papers by 23% more authors. The journals Bioinformatics and Nucleic Acid Research published the most papers with 810 and 409 papers respectively, and these were also the top two journals in 2008. The field continues to grow with more authors, journals, and papers contributing to the scientific discourse.

Analysis of topics, title keywords, and biological products point to several defining topics. The top 6 major MeSH topics were methods, metabolism, genetics, chemistry, computational biology (which in Mesh stands for Bioinformatics as well), and proteomics. In 2008, the most frequent topics were computational biology, genomics, proteomics, algorithms, proteins, and software. Further analysis of bioinformatics methods papers in 2010 (as identified by MeSH topics) reveals frequently occurring topics such as gene expression profiling, mass spectrometry, DNA sequence analysis, oligonucleotide array sequence analysis, and protein interaction mapping.

The top 10 terms appearing in the title of bioinformatics papers in 2010 were analysis, protein, gene, human, data, proteomic, expression, identification, and proteomics. The top 10 studied biological products included proteins, biological markers, bacterial proteins, messenger RNA, peptides, microRNAs, transcription factors, biological tumor markers, DNA, and ligands.

Ninety countries contributed papers falling within the criteria of the present survey. The top 10 contributing countries were the United States, Germany, China, United Kingdom, Japan, Canada, France, Italy, India, Spain, and the Netherlands. China, Germany, and Japan were the countries with the most papers. Among the four last authors, Ying Xu, Xia Li, Matthias Mann, and Satoru Miyano, contributed the most: 48 papers. The United States and China contributed the top 5 first authors with highest productivity: Vladimir Uversky, Meng Chen, Yijun Meng, Bin Xue, and Qing Yan, who contributed 26 papers.

## Informatics for Executing Complex Molecular Assays

***Next Generation Sequencing.*** The extremely rapid expansion of next-generation, high-throughput DNA sequencing technology (NGS) is arguably the most important scientific development of 2010 impacting the work of bioinformaticians. The availability of NGS at increasingly lower prices and larger data generation capability has led to its application to an extremely wide range of biological problems including most areas of basic biomedical and clinical translational research. From an informatics perspective, large numbers of laboratory and clinical scientists are empowered by NGS to generate extremely large data sets, such as multiple sets of paired tumor and healthy genomes at 30x coverage (100 Gb per genome) or shotgun meta-genomic data sets with >100 million 2x100 bp reads, which require the development of new analytical tools and methods as well as a substantial investment of time for data analysis by skilled informaticians. Fields such as cancer genomics, environmental and medical microbiology, and epidemiology have been dramatically revitalized and transformed by the availability of huge data sets generated by NGS. Clinical applications such as prenatal diagnosis for genetic abnormalities [2] and carrier testing for heritable diseases [3] is becoming feasible with very small and/or mixed samples as informatics methods improve the detection of true mutations against a background of normal cells and sequencing errors that produce false positives by using quality scores and local depth of coverage adjustments, such as in the SAMTools [2] and Picard toolkits (http://picard.sourceforge.net).

The application of NGS to human genomics has generated some landmark research results in 2010, which have had a profound impact on bioinformatics. Not only did bioinformatics methods and collaborators make essential contributions to these projects, but these fundamental discoveries in basic biology provide the foundation for the development of new bioinformatics methods. The 1000 Genomes Project published a progress report [4] that characterized 15 million sequence variants, 1 million insertion/deletion events, and 20,000 structural variants, primarily in protein coding regions, in approximately 700 individual human genomes. This data can be used to accurately characterize the mutational burden of the average individual, regional selection pressure on DNA sequences in and near protein coding genes, and the *de novo* mutation rate for humans. The 1000 Genomes Project public data set (http://www. 1000genomes.org ) has already become a standard part of data analysis pipelines for Genome-Wide Association-Studies (GWAS) of inherited disease risk alleles and cancer genomic screens for somatic mutations.

Informatics for RNA-seq (transcriptome sequencing) improved in several ways in 2010. Improvements in sample preparation protocols developed by several NGS vendors allowed for much greater and more representative sampling of non-protein-coding RNA molecules in RNA-seq data [5]. Methods such as nanoCAGE that combine novel sample preparation from small biological samples and informatics techniques have been developed to identify a much greater diversity of Transcription Start Sites for an expanded universe of functional transcripts across the genome [6].

***Informatics of De Novo Assembly.*** Unlike RNA-seq, ChIP-seq and some other next generation sequencing applications, there are many technical hurdles for *de novo* assembly using short reads sequence data, especially for those genomes with regions rich in repeats [7]. A new assembly algorithm developed by Gnerre and his colleagues [8] has made a substantial breakthrough in the field. The new *de novo* assembly algorithm, ALLPATHS-LG, created genome assemblies of human and mouse with good accuracy, short-range contiguity, long-range connectivity and nearly complete coverage of the genome.

In contrast to the previous generation of short reads, de novo assembly tools such as SOAPdenovo [9, 10], the ALLPATHS-LG algorithm uses approximately 100bps short reads in an average 45X coverage for large mammalian genomes (human and mouse) and yields much better de novo assembly results in terms of N50 size of both contig and scaffold. The contiguity is about 4-fold longer and the connectivity is about 25-fold longer than obtained with SOAPdenovo. Regarding the coverage of reference genomes, the assembly from ALLPATHS-LG contains 91.1% of the reference genome and 95.1 of the exonic bases whereas the assembly from SOAPdenovo covered only 74.3% of the genome and 81.2% of exonic bases for a human genome assembly. Similar results were obtained for a mouse genome assembly.

***Informatics of Microbiomics and Metagenomics.*** Microbiomic research has thrived with the expansion of sequencing capacity - a survey paper lists over 5800 sequencing projects underway with many close to completion [11]. The Human microbiome project has generated a number of subprojects and affiliated projects that focus on the microbiota of specific body sites; these started releasing independent databases for their domains [12-14]. The informatics analysis for microbiomics and metagenomics often require an assembly of multiple tools, which lead informatics groups to start releasing their novel methods and compendia of tools in the form of prebuilt virtual machine images, some of which are deployable on cloud clusters (e.g., http://clovr.org/; [15-17]).

***Proteomics Informatics.*** The developments in proteomics informatics during 2010 were dominated by significant refinements of methods, making protein identification and quantitation using mass spectrometric data more robust and more widely applicable.

Protein identification is usually done by searching collections of protein sequences from mass spectrometric data, and is, therefore highly dependent on the availability of protein sequence information. With the drop in cost and increase in speed of DNA sequencing, it is now possible to sequence the transcriptome of an organism within the scope of a proteomics project even if there are not sufficient sequences already available. This has made protein identification more widely useful, and will decrease the need to resort to the more challenging de novo sequencing strategies.

There are now mass spectrometric observations of a large fraction of all human genes publicly available, and they can be used for validation of results and experimental planning. For example, there are now more than 250 million high-quality tandem mass spectra in GPMDB (gpmdb.thegpm.org). This trend of ever more data becoming publicly availably will continue at a faster rate as journals and funding agencies increasingly require the submission of raw data. In the beginning of 2010 the journal, Molecular and Cellular Proteomics, started requiring the submission of all raw mass spectral data prior to publication of a manuscript. Also, a gene centric Human Proteome Project was announced this year [18], and it promises to generate vast amounts of data. These data can also be used for protein identification by searching spectral libraries [19, 20] which gives greater sensitivity because the intensity information in the tandem mass spectra can be utilized and the search is restricted to peptides that are commonly observed. As the coverage of spectral libraries increases, we will see a trend where protein identification is done less by searching protein sequence collections and more by searching spectral libraries.

Proteomics has moved from being qualitative to quantitative, and it is no longer possible to publish simply a list of proteins identified in a sample. Protein quantitation can be done using the mass spectrometric peak intensity of peptides or their fragments, or by counting the number of times a peptide is identified. The software available is still rapidly improving [21, 22], but has now matured to the degree that it is widely used.

## Informatics for Linking Molecular Data to Biological and Medical Phenotypes for Next-generation Diagnostics and Personalized Medicine

***Advances in clinical-grade molecular signatures/markers.*** A most important development and ongoing trend in translational bioinformatics is the proliferation of FDA-approved molecular profiling modalities for diagnosis and personalization of treatments. Such a “theranostic” is Oncotype Dx from Genomic Health which has proven not only clinically useful but also very successful financially. Specifically, as of September 30, 2010, more than 10,000 physicians in over 55 countries had ordered more than 175,000 Oncotype DX tests (http://www.genomichealth.com/). This success is fueling investments in the creation, validation and marketing of many more clinically useful molecular signatures. Since the first such molecular signature MammaPrint from Agendia (http://www.agendia.com) came onto the market in 2007, several dozen new modalities have been marketed and are offered/used for clinical practice, see Table 1 for major examples.

Bioinformatics methods and analyses are very important for creating and validating such molecular profiles, both for executing the requisite high throughput molecular assays as well for creating models that diagnose or predict clinical phenotypes on the basis of complex molecular marker patterns. Although it is not the purpose of this review (nor is it easy) to make accurate estimates, we informally estimate, based on public announcements from companies such as the ones listed in Table 1, that dozens to hundreds more are currently under development and in testing phase. We base these estimates on the fact that most companies that have marketed or are testing such products in trials, have announced pipelines with more than five new products per company and that these product announcements typically underestimate the number of the new products under development.

As we stated earlier, molecular profiling is the product of convergence of high-throughput molecular assays and sophisticated algorithms for linking clinical phenotypes to the molecular information. Such algorithms perform variable selection and transformation/construction, classification and regression, estimation of generalization error of the produced models, conversion and optimization of “discovery” models into “production“”models etc. Thus quality Bioinformatics methods and expertise are more than assistive in this translational enterprise: they have become essential. Along these lines we note the following developments and trends.

***Continued lack of best practices and non-rigorous method development cycles.*** Numerous methods for linking phenotypes to genetic, epigenetic, genomic and post genomic information involves reducing dimensionality via variable selection and classification methods. Many additional methods deal with model selection and error estimation issues, with combining algorithms into robust data analysis protocols and with software that implements such protocols (e.g., GEMS system [23] or FAST-AIMS system [24]). Even fewer studies deal with direct and comprehensive comparisons of methods (e.g., benchmarking studies: [25-29]). Furthermore, in our experience, many journals (e.g., the Bioinformatics journal) have not been recently interested in benchmarking of existing methods as much as in publishing new methods. This unfortunately applies to problem classes for which hundreds of methods were previously published but with limited validation. A disconcerting pattern that continued in 2010 therefore is that new methods are introduced with often minimal theoretical guarantees of performance, and limited empirical evidence for their comparative advantages and disadvantages to previous methodologies. As a result, informatics methods are often assumed to be interchangeable by the biological community [30] and few informatics consulting services and collaborative science efforts in the industry or academia are following best practices or evidence-driven-oriented approaches.

***Implications of problems with Bioinformatics methods: lessons from the Duke experience.*** A series of developments in the last year (Cancer Letter; Volume 36, No. 28, July 23, 2010, available online from http://cancerletter.com/downloads/20100803_10) that has brought into sharp focus the desirability of an evidence-driven approach is the discontinuation at Duke University of three

**Table 1** Examples of recent clinical-grade molecular profiles for diagnosis and personalized medicine

| Company | Product name | Disease/phenotype | Purpose | Website |
|---|---|---|---|---|
| Agendia | MammaPrint | Breast cancer | Risk assessment for the recurrence of distant metastasis in a breast cancer patient. | http://usa.agendia.com/en/mammaprint.html |
| Agendia | TargetPrint | Breast cancer | Quantitative determination of the expression level of estrogen receptor, progesteron receptor and HER2 genes. *This product is supplemental to MammaPrint.* | http://usa.agendia.com/en/targetprint.html |
| Agendia | CupPrint | Cancer | Determination of the origin of the primary tumor. | http://row.agendia.com/en/cupprint.html |
| University Genomics | Breast Bioclassifier | Breast cancer | Classification of ER-positive and ER-negative breast cancers into expression-based subtypes that more accurately predict patient outcome. | http://www.bioclassifier.com |
| Clarient | *Insight Dx Breast Cancer Profile* (formely *GeneRx Breast Cancer Profile* by *Prediction Sciences*) | Breast cancer | Prediction of disease recurrence risk. | http://www.clarientinc.com/default.aspx?tabid=340 |
| Clarient | Prostate Gene Expression Profile | Prostate cancer | Diagnosis of grade 3 or higher prostate cancer. | http://www.clarientinc.com/Default.aspx?tabid=403 |
| Prediction Sciences | RapidResponse c-Fn Test | Stroke | Identification of the patients that are safe to receive tPA and those at high risk for HT, to help guide the physician's treatment decision. | http://www.predict.net/Prediction_Sciences/Stroke.html |
| Genomic Health | OncotypeDx | Breast cancer | Individualized prediction of chemotherapy benefit and 10-year distant recurrence to inform adjuvant treatment decisions in certain women with early-stage breast cancer. | http://www.oncotypedx.com/ |
| *bioTheranostics* (previously *AviaraDx*) | CancerTYPE ID | Cancer | Classification of 39 types of cancer. | http://www.aviaradx.com/cTYPE/cType_overview.html |
| *bioTheranostics* (previously *AviaraDx*) | Breast Cancer Index | Breast cancer | Risk assessment and identification of patients likely to benefit from endocrine therapy, and whose tumors are likely to be sensitive or resistant to chemotherapy. | http://www.aviaradx.com/MGI_combo/combo_overview.html |
| Applied Genomics | MammaStrat | Breast cancer | Risk assessment of cancer recurrence. | http://www.applied-genomics.com/mammostrat.html |
| Applied Genomics | PulmoType | Non-small cell lung cancer | Classification of non-small cell lung cancer into adenocarcinoma versus squamous cell carcinoma subtypes. | http://www.applied-genomics.com/pulmotype.html |
| Applied Genomics | PulmoStrat | Lung cancer | Assessment of an individual's risk of lung cancer recurrence following surgery for helping with adjuvant therapy decisions. | http://www.applied-genomics.com/pulmostrat.html |
| Correlogic | OvaCheck | Ovarian cancer | Early detection of epithelial ovarian cancer. | http://www.correlogic.com/research-areas/ovarian-cancer.php |
| LabCorp | OvaSure | Ovarian cancer | Assessment of the presence of early stage ovarian cancer in high-risk women. | http://www.nytimes.com/2008/08/26/health/26ovar.html |
| Veridex | GeneSearch BLN Assay | Breast cancer | Determination of whether breast cancer has spread to the lymph nodes. | http://www.veridex.com/GeneSearch/GeneSearchHCP.aspx |
| Power3 | BC-SeraPro | Breast cancer | Differentiation between breast cancer patients and control subjects. | http://www.power3medical.com/products/bcserapro.aspx?level=0 |

clinical trials of molecular profile based modalities for cancer personalized treatment, stopping of funding for 3 related grants, paper retractions, a patent denial, a 3-year internal investigation at Duke and most recently resignation of a prominent Principal Investigator from the Duke faculty (http://dukechronicle.com/article/anil-potti-duke-cancer-researcher-accused-misconduct-resigns). Legal repercussions might also arise in the future. The underlying cause of all these events was improper bioinformatics and computational data analytics that compromised the validity of the molecular profiling models [31, 32].

***Advances in benchmarking studies.*** Reference [33] provides a comparison of microarray-based survival analysis algorithms and concludes that methods using coefficient shrinkage or linear combinations of the gene expression values yield better performance than simple variable selection methods; with ridge regression showing the overall best performance. Reference [27] conducted a comprehensive comparison of Random Forests with SVMs, two of the most prominent classifiers for gene expression microarray data and found that SVMs are superior. These results combined with those from an older evaluation [26] show that SVMs constitute a classifier of choice for molecular profiling. Currently the landscape is not as clear for variable selection. Recent developments from computer science/machine learning [29] point in the direction of recently introduced sound and scalable Markov Boundary methods which may have advantages over other methodologies. However a large-scale conclusive study for genomics data is currently lacking.

***Multivariate vs Univariate analyses for GWAS studies.*** The year 2010 continued a trend seen in 2009, witnessing the emergence of multivariate analysis of GWAS data [34-37]. Although more than 1,000 GWAS studies have been completed and corresponding datasets made available [38] (http://www.genome.gov/gwastudies/) the dominant analysis paradigm was univariate (one SNP at a time). In many cases univariate signals are weak and sometimes non-reproducible. Non-reproducibility can be attributed to underpowered studies (relative to the weak univariate signals), with too low of an a-priori probability that any single SNP will be associated with a phenotype, to batch effects [39], and to insufficient control of multiple comparisons (a problem that has been much overcome in recent years). The recently emerging multivariate analyses for GWAS data, show that many datasets with very small univariate effects have stronger multivariate signal. This gives hope that such data may be usable in the future for clinical prognostic and predictive purposes (comparable to the role that microarray data play for clinically applicable molecular profiles as explained in the "Advances in clinical-grade molecular signatures/markers" section.

***Advances in the analysis of signature and marker multiplicity.*** Another notable methodological development in 2010 involves issues of biomarker and molecular signature multiplicity (also known as the "Rashomon effect") [40, 41]. This phenomenon is very prominent in microarray and other omics data, and creates serious interpretability problems for any set of biomarkers or signatures that are effectively equivalent with a multitude of markers and signatures that fit the data equally well. Paper [42] provides for the first time algorithms that extract *all equivalent signatures and markers that satisfy strict optimality criteria.* In addition the paper shows that multiplicity is not necessarily the result of small sample size or deterministic relationships, it does not undermine reproducibility, and is not necessarily the result of biological pathway functional redundancies, although these factors and several others may contribute to the phenomenon. While at the present this phenomenon is not entirely understood [42] describes powerful computational tools for extracting all equivalence class members and for understanding its causes. The algorithms introduced in the above paper were shown to also offer significant reproducibility advantages over previous methods.

***Ability to predict clinical outcomes is real but limited compared to diagnosis.*** The ability to differentiate between various phenotypes using molecular information and computer models of that information has been firmly established. The ability to predict clinical outcomes was strongly challenged by the work of Michiels et al. [30]. Work by Aliferis et al. [28] however revealed that the doubts cast by Michiels et al. were due to an underpowered analysis protocol and deficiencies in the statistical testing employed. The results in [28] strengthen the belief in the technical feasibility of developing personalized medicine computer models even with modest sample sizes. They also show that thinking about power in omics studies need be accomplished at the *composite (multivariate) signal level and not one variable at a time as is often done.* This work finally posits that *efficiency* is of paramount importance and the usual "call to arms" in the literature for much larger samples for omics studies needs to be carefully balanced against scarce research resource allocation.

At the same time we observe that prognostic and predictive computational modeling tasks are much harder than diagnostic tasks. Typical signal strengths in the domain of cancer, as, for example, measured by AUC, are 90% or better for diagnosis but only 75% in the best case for prognosis/outcome prediction. It is not yet understood what are the reasons for this discrepancy and it is outside the scope of the present review to cover all plausible reasons. We do note however that this is an area where progress is badly needed, and if successful, will translate to substantial clinical benefits.

## Informatics for Biological Knowledge Discovery

***Advances catalyzed by new HT assay developments.*** The applications of the proteomic methods mentioned earlier for protein identification and quantitation vary widely. A few examples include: elucidation of how genetic networks are rewired in response to DNA damage [43], correlation between gene copy number and protein amount in cancer cells [44], understanding of cystic fibrosis [45], and investigation of virus-host protein interactions during infection [46].

The Cancer Genome Atlas Network published papers analyzing the genomic sequences of large patient cohorts for glioblastoma [47, 48] and acute myeloid leukemia [49]. In glioblastoma, a variety of mutations have been discovered in specific genes such as EGFR, NF1, and PDGFRA/IDH1 that are associated with specific cancer subtypes, which have previously been characterized by gene expression signatures and correlated with responses to therapy. DNMT3A mutations are highly recurrent in patients with *de novo* AML with an intermediate-risk cytogenetic profile and are independently associated with a poor outcome. These findings validate the general approach of tumor sequencing and variant discovery in order to identify concordant mutations in key genes or pathways that lead to malignant transformation and/or more aggressive disease. Many more cancer sequencing projects are underway due to these promising early findings.

New sequencing informatics methods for 2010 were primarily incremental improvements in key algorithms in areas such as *de novo* genome assembly, alignment of short reads to reference genomes, sequence variant detection, quantification of gene expression, detection of alternative or novel splice isoforms, identification of epigenomic changes, metagenomic taxonomic identification and comparisons of bio-samples by species abundance and diversity using primarily unsupervised learning techniques.

In 2010, the computational analysis of ChIP-seq data became more precise, allowing for studies that focus on tissue specific differences, response to experimental treatment, clinical changes, and evolutionary conservation of changes in transcription factor binding [50], nucleosome position, and histone methylation. New analysis methods also allowed for the identification of changes in the shape or extent of DNA regions affected by histone modification. Additional software packages have been developed for ChIP-seq that improve statistical methods [51] sensitivity of transcription factor binding site detection [52] and annotation of large data sets [53, 54].

***Sequence variant discovery***: Massive parallel sequencing technologies substantially enhance the opportunities for study of DNA sequence variation, particularly the identification of variants that are associated with human disease. A number of open source tools have been developed to analyze alignments of NGS reads to a reference genome and detect variants. Several widely used open source tools have been developed in the past few years such as SAMTools [55] and VarScan [56]. The performance of these tools is improving. The latest development in variant detection is the comprehensive package GATK developed in 2010 [57]. GATK was designed to simplify the process of developing efficient, robust tools for working with NGS data in large binary encoded sequence file using methods optimized for common modes of data access that emphasize correctness, stability, CPU and memory efficiency, and enable distributed and shared memory parallelization. GATK currently supports in a single integrated framework Illumina, SOLiD, 454, Complete Genomics, and Sanger sequencer data. Using this framework, a number of widely-used tools have been developed and released such as base quality score recalibration, local realignment around indels, multi-sample SNP and indel callers, as well as read and variation QC tools. These tools are now integrated into the 1000 Genomes Project, The Cancer Genome Atlas, Broad's production sequencing pipeline, as well as at those at many other sequencing centers and individual labs with sequencing machines. Compared to other tools for variants calling, GATK aims to achieve better precision and fewer false positives through the use of base quality score recalibration and local realignment around indels, while still maintaining sensitivity.

***Evidence against the common variant hypothesis.*** The works by Goldstein et al. [58, 59] provided evidence that what is observed as genotype-phenotype associations is in fact a manifestation of different loci that are not necessarily captured by the common SNPs typically studied in GWASs. Furthermore, this work shows that the true causal variants may be located anywhere in the genome and hence may be in genomic locations *near or far away* from the SNPs that exhibit strong predictive signals for the phenotype being studied. These results necessitate the combined examination of both common and rare variants or at least the development of discovery methods that, under reasonable assumptions, can uncover rare variant causes of disease from common variant information.

***Improvements in causal graph-based variable selection and in understanding the relationship between causation and predictivity as a function of analysis method.*** When biologists seek to discover the drivers and not the bystanders of phenotypes, standard and common variable selection procedures can be highly misleading. This is the main result of a recent massive evaluation of variable selection methods, some utilizing causal induction and some strictly associative, that also reveals that causal induction of the local direct (relative to the measured variables) causes and effects of the response variable (i.e., phenotype) yields a most predictive marker set that has also maximum parsimony and furthermore is locally causal consistent with the function that generates the data [29]. This work highlights that the extensive and common use of clustering, univariate association as well as more sophisticated variable selection methods such as SVM-based, random forest-based and penalized regression-based can be highly misleading in terms of mechanistic determinants of the phenotype. These results also provide a "silver lining" in that formal causal graph methods are promising in that they provide better theoretical guarantees and strong empirical performance in the studied experiments, suggesting the value of further research on these approaches.

***Quality of pathway reverse engineering methods.*** A staple of biological research is methods for de novo reconstruction of pathways from observational or quasi-experimental data (i.e., where only a few out of many variables have been manipulated). A recent discovery competition http://wiki.c2b2.columbia.edu/dream/results/DREAM5/?c=4_1 identified Random Forests as a winning pathway induction method. A much more comprehensive and rigorous benchmark of local pathway reconstruction from case-control data [25] examined all core methods for the task and revealed that some methods need to be substantially improved upon (e.g., clustering, relevance networks), while others should be used routinely (several causal graph-based methods). The paper also demonstrated that several univariate methods provide a "gatekeeper" performance threshold that should be applied when method developers assess the performance of their novel multivariate algorithms. This work also highlighted the effect of loss function (i.e., corresponding to researcher experimental resource constraints) on the choice of best discovery methodology.

***Methods for minimizing the number of experiments and for automated or semi-automated discovery.*** In the past years several systems and methods have been introduced for selecting experiments automatically and, in some cases, also automatically conducting the experiments. The main work includes: *ILVS, LIM, GEEVE* [60-66], the Tong and Koller method [67], the Murphy method [68], *LLC* [69], the Pournara and Wernisch method [70], *ALCBN* [71], the He and Geng method [72], *Adam* [73-76], *GenePath* [77-80], the Ideker *et al. method* [81], *MEED* [82], the Tegner *et al. method* [83], and the Steinke *et al. method* [84]. Research in *strategic experimentation for causal discovery* can be valuable in providing theoretical bounds on the number of experiments needed to unravel causal relations within a set of variables [85-87] and proposes strategies to select a set of variables for an experiment, aiming at minimizing the total number of experiments [88, 89]

***Methods for causal orientation of pairs of variables from observational data.*** We close this section by mentioning a very exciting new development in computational causal discovery which involves techniques to identify the direction of causality among pairs of variables when it is known that they are causally related, but where the direction is unknown. This problem was deemed to be unsolvable previously without experimentation but it is addressed by the new approaches by exploiting the asymmetry of information when going from cause-to-effect and vice-versa [90-94].

## Conclusions

We conclude this survey report by summarizing „Big Picture“ issues such as trends, major insights gained during the previous year and ongoing challenges and open problems that need be solved.

1. *Bioinformatics in the post-Human genome project era continues to be a key ingredient and a major driver in the core progress of Basic Biology and of Translational Medicine.*
2. *Several new technologies that enable unprecedented accuracy in molecular studies of all type of disease models have emerged, most notably Next Generation Sequencing.* All these technologies depend heavily on bioinformatics to operate (what we have referred to as *Informatics for executing complex molecular assays*). Making sense of the resulting measurements either at the level of biological disease mechanism or at the level of diagnosis and outcome prediction critically depends on bioinformatics as well (we called these type of analyses *Informatics for biological knowledge discovery* and *Informatics for knowledge discovery related to managing clinical phenotypes*, respectively).
3. *The prototypical molecular translational methodology in the literature is molecular profiling. Clinical-grade molecular signatures/ markers are proliferating at an exploding rate.* Critical for their development is sound computational data analysis.
4. *Developing best practices in all aspects of bioinformatics analysis is greatly needed and has severe implications for translational and basic science.* While many new methods for all aspects of such analysis exist and continue to be developed, less emphasis is being placed, unfortunately, on their theoretical understanding and empirical comparison. The events surrounding the breakdown of several related projects at Duke, suggest that deficiencies in informatics and computational data analytic methods can have catastrophic consequences for otherwise solid and important scientific endeavors.
5. *A large number of studies deals with introducing new methods but only a tiny fraction addresses their theoretical and empirical evaluation. The few systematic and comprehensive benchmarking studies of old and new methods that have been conducted in the last few years invariably produced important findings.* We discussed a few of those studies which suggest that several methods in widespread use need be retired (or imporved) and other methods that are less used are more robust and effective and deserve to be used more widely.
6. *Several exciting new methods have appeared that open new directions for progress in the field.* As examples we mention new ways to deal with multiplicity, improved and advanced causal graph methods, and new theory and algorithms for minimizing the number of experiments needed to uncover complex biological pathways. *At the same time the range of open problems of great significance is extensive.* They include: what causes multiplicity and how can it be reduced? Why predictive signals in clinical outcome studies are so much weaker than those in diagnostic studies ,and how they can be improved? How can destructive batch effects be prevented, detected and corrected?

We look forward to next year's developments with great eagerness, expecting many of these questions to have been addressed, leading, no doubt, to unanticipated advances as well.

**Acknowledgments**

Supported in part by grant 1 UL1 RR029893 from the National Center for Research Resources, National Institutes of Health.

## References

1. Kohane I: Ten thousand views of bioinformatics: a bibliome perspective. Yearb Med Inform 2009:113-6.
2. Lo YM, Chan KC, Sun H, Chen EZ, Jiang P, Lun FM, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med 2010,2:61ra91.
3. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 2011,;3:65ra4.
4. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature 2010,;467:1061-73.
5. Kapranov P, St LG, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, et al. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is ‚dark matter' un-annotated RNA. BMC Biol 2010;8:149.
6. Plessy C, Bertin N, Takahashi H, Simone R, Salimullah M, Lassmann T, et al. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. Nat Methods 2010;7:528-34.
7. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods 2011;8:61-65.
8. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 2011;108:1513-8.
9. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 2010;20:265-72.
10. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of the giant panda genome. Nature 2010;463:311-7.
11. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, et al. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 2010;38:D346-54.
12. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, et al. A catalog of reference genomes from the human microbiome. Science 2010;328:994-9.
13. Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, Hankeln W, et al. Megx.net: integrated database resource for marine ecological genomics. Nucleic Acids Res 2010;38:D391-5.
14. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford) 2010;2010:baq013.
15. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7:335-6.
16. Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, Methe BA, et al. METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. Bioinformatics 2010;26:2631-2.
17. Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb Protoc 2010;2010:db.
18. A gene-centric human proteome project: HUPO—the Human Proteome organization. Mol Cell Proteomics 2010; 9:427-9.
19. Fenyo D, Eriksson J, Beavis R. Mass spectrometric protein identification using the global proteome machine. Methods Mol Biol 2010;673:189-202.
20. Lam H, Aebersold R. Using spectral libraries for peptide identification from tandem mass spectrometry (MS/MS) data. Curr Protoc Protein Sci 2010; Chapter 25:Unit 25.5.
21. Cox J, Matic I, Hilger M, Nagaraj N, Selbach M, Olsen JV, et al. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. Nat Protoc 2009;4:698-705.
22. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 2010;26:966-8.
23. Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis CF. GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. Int J Med Inform 2005;74:491-503.
24. Fananapazir N, Statnikov A, Aliferis CF. The FAST-AIMS Clinical Mass Spectrometry Analysis System. Adv Bioinformatics 2009:598241.
25. Narendra V, Lytkin NI, Aliferis CF, Statnikov A. A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. Genomics 2011;97:7-18.
26. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multi-category classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005;21:631-43.
27. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 2008;9:319.
28. Aliferis CF, Statnikov A, Tsamardinos I, Schildcrout JS, Shepherd BE, Harrell FE. Factors Influencing the Statistical Power of Complex Data Analysis Protocols for Molecular Signature Development from Microarray Data. PLoS ONE 2009;4:e4922.
29. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. Journal of Machine Learning Research 2010;11:171-234.
30. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488-92.
31. Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. Annals of Applied Statistics 2009;3:1309-34.
32. Coombes KR, Wang J, Baggerly KA. Microarrays: retracing steps. Nat Med 2007;13:1276-7.
33. Bovelstad HM, Nygard S, Storvold HL, Aldrin M, Borgan O, Frigessi A, et al. Predicting survival from microarray data—a comparative study. Bioinformatics 2007;23:2080-7.
34. Jiang X, Neapolitan RE, Barmada M, Visweswaran S, Cooper GF. A Fast Algorithm for Learning Epistatic Genomic Relationships. AMIA 2010 Annual Symposium Proceedings 2010;:341-5.
35. Cooper GF, Hennings-Yeomans P, Visweswaran S, Barmada M. An Efficient Bayesian Method for Predicting Clinical Outcomes from Genome-Wide Data. AMIA 2010 Annual Symposium Proceedings 2010;:127-31.
36. Alekseyenko AV, Lytkin NI, Ai J, Ding B, Padyukov L, Aliferis CF, et al. Causal Graph-Based Analysis of Genome-Wide Association Data in Rheumatoid Arthritis.: CHIBI Technical Report 2010-2, New York University Langone Medical Center; 2010.
37. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet 2009;5:e1000678.
38. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009;106:9362-7.
39. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010;11:733-9.
40. Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 2003;19:1484-91.
41. Azuaje F, Dopazo J. Data analysis and visualization in genomics and proteomics. Hoboken, NJ: John Wiley; 2005.
42. Statnikov A, Aliferis CF: Analysis and Computational Dissection of Molecular Signature Multiplicity. PLoS Comput Biol 2010; 6:e1000790.
43. Bandyopadhyay S, Mehta M, Kuo D, Sung MK, Chuang R, Jaehnig EJ, et al. Rewiring of genetic networks in response to DNA damage. Science 2010;330:1385-9.
44. Geiger T, Cox J, Mann M. Proteomic changes resulting from gene copy number variations in cancer cells. PLoS Genet 2010;6.
45. Koulov AV, Lapointe P, Lu B, Razvi A, Coppinger J, Dong MQ, et al. Biological and structural basis for Aha1 regulation of Hsp90 ATPase activity in maintaining proteostasis in the human disease cystic fibrosis. Mol Biol Cell 2010;21:871-84.
46. Terhune SS, Moorman NJ, Cristea IM, Savaryn JP, Cuevas-Bennett C, Rout MP, et al. Human cytomegalovirus UL29/28 protein interacts with components of the NuRD complex which promote accumulation of immediate-early RNA. PLoS Pathog 2010;6:e1000965.

47. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010;17:98-110.
48. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS One 2010;5:e8918.
49. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, et al. DNMT3A mutations in acute myeloid leukemia. N Engl J Med 2010;363:2424-33.
50. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 2010;328:1036-40.
51. Zhang X, Robertson G, Krzywinski M, Ning K, Droit A, Jones S, et al. PICS: Probabilistic Inference for ChIP-seq. Biometrics 2010.
52. Hu M, Yu J, Taylor JM, Chinnaiyan AM, Qin ZS. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. Nucleic Acids Res 2010;38:2154-67.
53. Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 2010;11:237.
54. Ye T, Krebs AR, Choukrallah MA, Keime C, Plewniak F, Davidson I, et al. seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res 2011 Mar;39(6):e35.
55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.
56. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 2009;25:2283-5.
57. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.
58. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biol 2010; 8:e1000294.
59. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 2010;11:415-25.
60. Cooper GF, Yoo C. Causal Discovery from a Mixture of Experimental and Observational Data. Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-99) 1999;:116-25.
61. Yoo C, Thorsson V, Cooper GF. Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. Proceedings of the 2002 Pacific Symposium on Biocomputing 2002;:498-509.
62. Yoo C, Cooper GF. Discovery of gene-regulation pathways using local causal search. Proc AMIA Symp 2002;:914-8.
63. Yoo C, Brilz EM. The five-gene-network data analysis with local causal discovery algorithm using causal Bayesian networks. Ann N Y Acad Sci 2009;1158:93-101.
64. Yoo C, Cooper GF. A computer-based microarray experiment design-system for gene-regulation pathway discovery. AMIA Annu Symp Proc 2003;:733-7.
65. Yoo C, Cooper GF. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. Artif Intell Med 2004;31:169-82.
66. Yoo C, Cooper GF, Schmidt M. A control study to evaluate a computer-based microarray experiment design recommendation system for gene-regulation pathways discovery. J Biomed Inform 2006;39:126-46.
67. Tong S, Koller D. Active learning for structure in Bayesian networks. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001) 2001;17:863-9.
68. Murphy KP. Active learning of causal Bayes net structure. Technical Report, University of California, Berkeley; 2001.
69. Eberhardt F, Hoyer PO, Scheines R. Combining Experiments to Discover Linear Cyclic Models with Latent Variables. Journal of Machine Learning Research, Workshop and Conference Proceedings (AISTATS 2010) 2010;9:185-92.
70. Pournara I, Wernisch L. Reconstruction of gene networks using Bayesian learning and manipulation experiments. Bioinformatics 2004;20:2934-42.
71. Meganck S, Leray P, Manderick B. Learning Causal Bayesian Networks from Observations and Experiments: A Decision Theoretic Approach. Modeling Decisions in Artificial Intelligence, LNCS 2006:58-69.
72. He Y, Geng Z: Active learning of causal networks with intervention experiments and optimal designs. J Mach Learn Res 2008;9:2523-47.
73. King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, et al.The automation of science. Science 2009;324:85-9.
74. King RD, Whelan KE, Jones FM, Reiser PG, Bryant CH, Muggleton SH, et al. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 2004;427:247-52.
75. Sparkes A, Aubrey W, Byrne E, Clare A, Khan MN, Liakata M, et al. Towards Robot Scientists for autonomous scientific discovery. Autom Exp 2010;2:1.
76. Wolinsky H. I, scientist. Will robots at the bench leave scientists free to think? EMBO Rep 2007;8:720-2.
77. Demsar J, Zupan B, Bratko I, Kuspa A, Halter JA, Beck RJ, et al. GenePath: a computer program for genetic pathway discovery from mutant data. Stud Health Technol Inform 2001;84:956-9.
78. Juvan P, Demsar J, Shaulsky G, Zupan B. GenePath: from mutations to genetic networks and back. Nucleic Acids Res 2005;33:W749-52.
79. Zupan B, Bratko I, Demsar J, Juvan P, Curk T, Borstnik U, et al. GenePath: a system for inference of genetic networks and proposal of genetic experiments. Artif Intell Med 2003;29:107-30.
80. Zupan B, Demsar J, Bratko I, Juvan P, Halter JA, Kuspa A, et al. GenePath: a system for automated construction of genetic networks from mutant data. Bioinformatics 2003;19:383-9.
81. Ideker TE, Thorsson V, Karp RM. Discovery of regulatory interactions through perturbation: inference and experimental design. Pac Symp Biocomput 2000:305-16.
82. Szczurek E, Gat-Viks I, Tiuryn J, Vingron M. Elucidating regulatory mechanisms downstream of a signaling pathway using informative experiments. Mol Syst Biol 2009;5:287.
83. Tegner J, Yeung MK, Hasty J, Collins JJ. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proc Natl Acad Sci U S A 2003;100:5944-9.
84. Steinke F, Seeger M, Tsuda K. Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models. BMC Syst Biol 2007;1:51.
85. Eberhardt F, Glymour C, Scheines R. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI-2005) 2005;:178-83.
86. Eberhardt F, Glymour C, Scheines R. N-1 Experiments Suffice to Determine the Causal Relations Among N Variables. Innovations in Machine Learning: Theory And Applications; 2006. p. 97-112.
87. Eberhardt F, Scheines R. Interventions and Causal Inference. Philosophy of Science 2007;74:981-95.
88. Eberhardt F. Almost Optimal Intervention Sets for Causal Discovery. Proceedings of 24th Conference in Uncertainty in Artificial Intelligence (UAI-2008) 2008;:161-8.
89. Eberhardt F. Causal Discovery as a Game. Journal of Machine Learning Research, Workshop and Conference Proceedings (NIPS 2008 causality workshop) 2010;6:87-96.
90. Peters J, Janzing D, Schölkopf B. Identifying Cause and Effect on Discrete Data using Additive Noise Models. Journal of Machine Learning Research, Workshop and Conference Proceedings (AISTATS 2010) 2010;9:597-604.
91. Daniusis P, Janzing D, Mooij J, Zscheischler J, Steudel B, Zhang K, et al. Inferring deterministic causal relations. Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI-2010) 2010;:143-50.
92. Hoyer PO, Janzing D, Mooij J, Peters J, Schölkopf B. Nonlinear causal discovery with additive noise models. Advances in Neural Information Processing Systems 2009;21:689-96.
93. Janzing D, Sun X, Schölkopf B. Distinguishing Cause and Effect via Second Order Exponential Models.: arXiv:0910.5561v1 [stat.ML]; 2009.
94. Zhang K, Hyvärinen A. Distinguishing causes from effects using nonlinear acyclic causal models. Journal of Machine Learning Research, Workshop and Conference Proceedings (NIPS 2008 causality workshop) 2008;6:157-64.

**Correspondence to:**
Constantin Aliferis MD, PhD
Center for Health Informatics and Bioinformatics, New York University
227 East 30th Street
New York, NY 10016, USA
Tel.: +1 212 263 5281
Fax: +1 615 469 3516
E-mail: constantin.aliferis@nyumc.org