# 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.

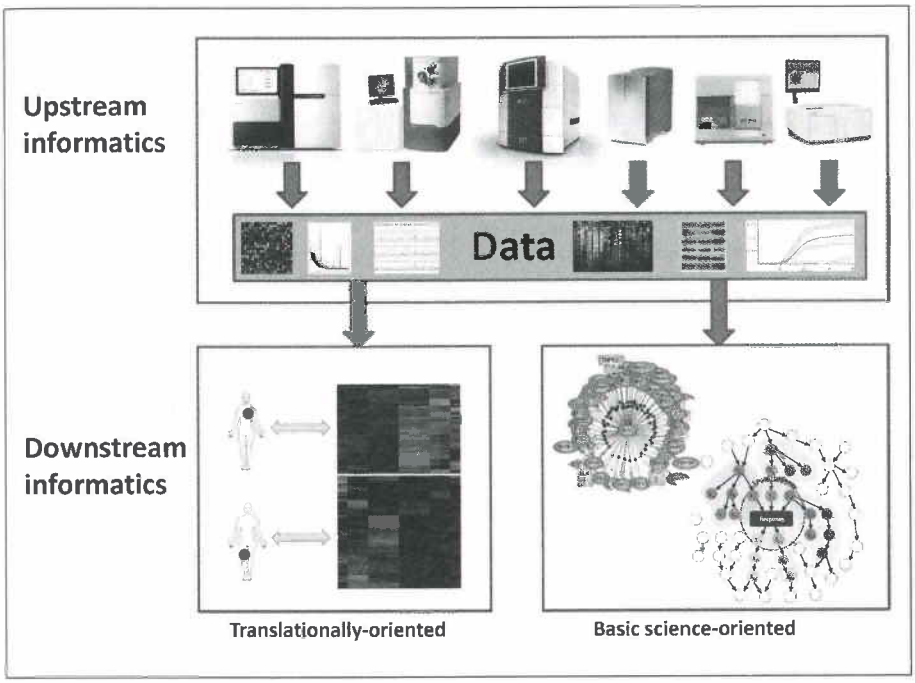


Fig. 1 Conceptualization of Bioinformatics field used in present review

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 bio-

logical 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 pre-built 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 mil-

lion 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 Ran-