SIGBio ACM

Bioinformatics, Computational Biology
and Biomedical Informatics

SIGBio Record

Newsletter of the SIGBio
ACM Special Interest Group

Volume 4, Issue 2, July 2014 ISSN 2159–1210

Notice to Contributing Authors to SIG Newsletters

Editor in Chief’s Notes:

This issue presents the following contributions:
A research paper from Tianle Ma; a profile of Yang Zhang by Amarda Shehu; a call for paper

We thank contributors for this issue and hope that readers will find interesting references to their work in Bioinformatics and Health Informatics area.

Pietro Hiram Guzzi
Young-Rae Cho
Pierangelo Veltri

Integrative and Interdisciplinary Challenges in Translational Bioinformatics

Tianle Ma
Department of Computer Science and Engineering,
State University of New York at Buffalo,
Buffalo, NY, 14260, USA
tianlema@buffalo.edu

ABSTRACT

Translational bioinformatics (TBI) is an emerging interdisciplinary field, which aims to bridge the gap between molecular world and clinical world. Translational bioinformatics employs data mining and machine learning techniques to analyze increasingly massive biomedical data and generate knowledge for clinical applications. One of the major challenges in TBI is to integrate multi-dimensional heterogeneous biomedical information sources in order to elucidate new biomedical knowledge. The integrative methodologies that are used to interpret these data require expertise in different disciplines, such as biology, medicine, mathematics, statistics and bioinformatics, and they pose great interdisciplinary challenges. Bioinformatics, system biology and network science together with knowledge engineering and reverse engineering have great potential to push TBI forward. In this paper, we introduce the background of TBI and the great variety of biomedical data, discuss the computational tools for integrative analyses, and summarize several crucial interdisciplinary challenges and future directions in TBI.

Categories and Subject Descriptors

H.2.8 **[Database Management]**: Database Applications—

Data Mining; J.3 **[Life and Medical Sciences]**: Health.

General Terms

Algorithms, Management, Theory.

Keywords

Translational Bioinformatics, Integrative Analysis, Systems Biology, Knowledge Engineering

INTRODUCTION

The accumulation of enormous quantities of molecular data and clinical data has led to the emergence of 'translational bioinformatics' (TBI), an inter-disciplinary field of study that focuses on linking molecular entities and clinical entities in order to better transform scientific discoveries into clinical applications. The American Medical Informatics Association defines translational bioinformatics as follows:

Translational bioinformatics is the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data, into proactive, preventive, predictive and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of TBI is newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders, including biomedical scientists, clinicians and patients.

The critical areas to be addressed in translational bioinformatics include the following categories [16]: 1) The management of multi-dimensional and heterogeneous molecular and clinical data sets; 2) The applications of knowledge-based systems and intelligent agents to enable high-throughput hypotheses generation and testing; 3) The facility of data-analytic pipelines in in-silicon research programs; and 4) The dissemination of data, information and knowledge generated during the translational cycle.

Though various informatics tools have been developed and great progress has been made in the abovementioned areas, there are still enormous challenges to understand the sophisticated mechanisms underlying complex biological systems and to optimize the transformation of the exploding biomedical data and information sources into significant clinical benefits. First, biological systems are inherently complex with numerous uncontrollable and unobservable variables, which pose extreme challenges to have a holistic view of all biological activities. To address this issue, systems biology, a biology-based inter-disciplinary field of study that focuses on complex interactions within biological systems, using a holistic approach to biological and biomedical research, has aroused great interests in recent years. Second, due to lack of standards and control of data quality in biological experiments as well as clinical trials, it's very difficult to comprehensively reuse the data collected in various contexts. Third, though increasingly massive biomedical data is linked together, usually through relational databases, e.g., NCBI, it's far more complex to effectively mine the data and extract useful information based on multiple heterogeneous datasets. Finally, though great efforts have been made, generally speaking, biomedical data, information, and knowledge are still scattered around or loosely connected. It becomes increasingly urgent to develop a unified framework which can automatically and comprehensively mine heterogeneous biomedical information sources. Thus, integrative analysis of biomedical data has been and is still intensively studied by researchers.

This paper aims to discuss several crucial challenges and outline a number of future directions in translational bioinformatics. The rest of the paper is organized as follows: we first introduce the various biomedical information sources and applications in Section 2. Then we discuss various integrative analysis approaches in translational bioinformatics in Section 3. Translational bioinformatics is inherently interdisciplinary. In

Section 4, we will discuss several interdisciplinary challenges and propose several corresponding crucial scientific inquiries in TBI. Finally we give a brief conclusion in Section 5.

EXPLODING BIOMEDICAL INFORMATION SOURCES

First, let's get a glimpse of the rich variety of biomedical data, including both molecular and clinical data. In the molecular world, we study molecules, including big molecules, such as DNAs, RNAs and proteins, and small molecules, such as amino acids, lipids, and sugars [1]. Instead of studying them in isolation, we want to figure out the molecular mechanisms of biological activities, including metabolites, cellular organization and communication, and various kinds of biological processes. This makes things quite complicated. Even though the advancements of biotechnology enable us to make various molecular measurements in an unprecedented way, we are far from getting a comprehensive understanding of the sophisticated mechanisms of molecular world in which trillions of molecules work together to fulfil numerous intertwined biological processes.

In the clinical world, we study diseases, drugs, patients, symptoms, clinical laboratory measurements, clinical images, and electronic health records (EHRs) [1]. However, the molecular world and clinical world are never separable. There are causality links between the two. In fact, translational bioinformatics (TBI) emerges to bring the gap between both molecular and clinical world. Great efforts have been made to integrate clinical and genomic data [21]. Table 1 lists a number of biomedical data and information sources.

Table 1. List of Biomedical Information Sources

Data Types	Contents	Techniques	Databases
Sequences	Nucleotide, peptide, etc.	Sequential analysis, typical software: BLAST	GenBank, RefSeq, PDB, etc.
Arrays	Gene expression	Statistics, Network analysis	GEO, Gene, BioProject
Metadata	Annotation and mapping information about the sequences	Text mining, integrative methods	dbSNP, dbVar, RefSeq, etc.
Unstructured Texts	Clinical records (EHRs), literature	Text Mining, NLP, integrative analysis	PubMed, (EHRs are usually not public available)
Structured Texts Knowledge	Gene ontologies, biological models, research finding, etc.	NLP, knowledge driven models	BioPortal (NCBO), BioProject, BioSystem, etc.

DNA and RNA: **1)** Nucleotide sequences and map data from the whole genomes of over 1000 organisms, public available from the International Nucleotide Sequence Database Collaboration; **2)** A collection of curated, non-redundant genomic DNA and transcript (RNA), which provides a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies, and comparative analyses, retrievable from Reference Sequence (RefSeq); **3)** Single nucleotide variations, micro-satellites, small-scale insertions and deletions, including population-specific frequency and genotype data, experimental conditions, molecular context, and mapping information for both neutral variations and clinical mutations, retrievable from Database of Short Genetic Variations (dbSNP); **4)** Large scale of genomic variation, including large insertions, deletions, translocations and inversions, public available from Database of Genomic Structural Variation (dbVar); and **5)** A collection of functional genomics, comparative genomics and genetic studies and their resulting datasets, retrievable from BioProject (formerly Genome Project). For a specific topic, there might be hundreds or thousands of studies with datasets public available. However, inconsistencies might exist [10] in different studies due to lack of standards, inconsistent annotations, experimental and environmental factors, the varied nature of many molecular measurements, such as gene expression profiles, etc. One of the major research inquires in TBI is how to integrate all these scatted research results and datasets, and generate a consistent data and knowledge repository.

Protein: **1)** Protein sequence records from a variety of sources, including PDB, RefSeq, PRF, etc; **2)** Known 3-dimensional protein structures; **3)** Sequence alignments and profiles representing protein domains conserved in molecular evolution; and **4)** Protein clusters, a collection of related protein sequences, including annotation information, domains and structures.

Gene and Gene Expression: **1)** Gene information, including nomenclature, sequences, chromosomal localization, variation details, expression reports, homologs, protein domain content, gene products and their attributes, associated markers, phenotypes, gene interactions, etc; **2)** Gene expression and molecular abundance profiles; and **3)** The relationships between human variations and observed health status with supporting evidence.

Genotype-Phenotype Mapping: **1)** The associations between genotype and phenotype from the results of studies which investigate the interaction of genotype and phenotype. These studies include Genome Wide Association Study (GWAS), medical resequencing, and so on. Typical public databases include dbGaP; **2)** The relationships between human variations and observed health status with supporting evidence, retrievable from ClinVar; and **3)** Human genes and genetic disorders, retrievable from OMIM.

General Clinical Data: **1)** Electronic Health Records (EHRs) could be used to deprive high quality phenotypic information [7]. One of TBI's inquiries is how to link EHRs with bio-bank in order to enhance clinical decision support systems. However, there are still a lot of challenges due to lack of standards, data qualities, technology barriers and privacy issues. In the past few years, Personal Health Records (PHRs) are emerging. Each individual's EHRs can be collected through one's life, which constitute Personal Health Records (PHRs). Both EHRs and PHRs contains temporal information, which can be used for temporal data mining [20]. PHRs will combine traditional EHRs and genomic data, which will maintain much more comprehensive information of individuals and will be of great value to multiple stakeholders in health care systems. However, we need to rebuild the existing IT infrastructure to support the transformation from EHRs to PHRs. Besides technology issues, privacy, ethic and legal concerns should also be taken into consideration; **2)** Clinical reports, drug responses, biomedical images and other lab measurements, could be useful for pharmacogenetics when linking to genomic information; and **3)** Results and datasets from numerous clinical trials could be grouped together so that novel information and knowledge could be discovered.

Biological Processes: Existing knowledge of high-level functions and utilities of the biological systems, such as molecular pathways, functional hierarchies, etc. Most of this information is manually curated and deprived from raw data described above.

Literature and Knowledge Repositories: **1)** Every year voluminous literature get published. For instance, more than 30, 000 articles on "biomedical" topic published in 2013 were retrievable from Web of Science. **2)** Great efforts have been made to link multiple data sources by using annotations and ontologies [15; 19; 22].

Almost all kinds of biomedical data related to a specific topic, for instance, a specific disease, has some relationships, either known or unknown to us. All the data are interconnected and coupled together. For example, once an identifier is assigned to the concept of a gene, multiple databases connected to that concept need to be updated, including databases on genes and gene products, pathways involving gene products, gene variations and corresponding phenotypes, literature, annotation and ontologies.

In general, biomedical information sources can be further divided into biomedical data and biomedical knowledge. For example, the nucleotide sequences and the 3-Dimensional structures of protein belong to the category of biomedical data, which contain the factual information as measurements or statistics used as a basis for reasoning, discussing, or calculation; gene ontologies and pathways can be classified as biomedical knowledge which is induced from biomedical data and formerly existing knowledge. There is great potential to combine data-driven and knowledge-driven approaches to generate new knowledge [2].

The major inquiry of TBI is how to incorporate various kinds of biomedical data and knowledge sources in a systematic way such that new knowledge can be generated which will directly benefit clinical practice.

INTEGRATIVE ANALYSES IN TRANSLATIONAL BIOINFORMATICS

With exploding biomedical data and knowledge sources, it becomes increasingly urgent to develop systematic frameworks to integrate multi-dimensional information sources. Numerous computational tools have been developed for integrative analyses in recent years. Yet they are far from perfection with various restrictions. Knowledge engineering in TBI is an emerging field which has aroused great interests in biomedical community.

Computational Tools for Integrative Analyses

There are three broad objectives of integrative analysis [14]: The first objective is to understand molecular behaviors, mechanisms and relationships between and within the different types of molecular structures, including associations between these and various phenotypes, such as clinical outcomes, pathways, interactions, etc. The second objective is to understand the taxonomy of diseases, thereby classifying individuals into latent classes of disease subtype. The third objective is to predict an outcome or phenotype for the prospective patients. [14] gives a detailed review of bioinformatics tools for integrative analyses in cancer, including sequential analysis, latent variable models, penalized likelihood, gene set analysis, pair-wise correlation methods, network-based analysis, Bayesian approaches, etc. Furthermore, [8] reviews cancer genomic software and the insights that have been gained from their applications. The bioinformatics tools and software for cancer can easily be extended to solve problems in other complex diseases.

Among many existing methods, integrative genomics is based on the fundamental principle that any biological mechanism builds upon multiple molecular phenomena, and only through the understanding of the interplay within and between different layers of genomic structures can one attempt to fully understand phenotypic traits [14]. Therefore, principles of integrative genomics are based on the study of molecular events at different levels and on the attempt to integrate their effects in a functional or causal framework. To infer causal relationships instead of mere statistical correlations and associations is the future direction of integrative analysis.

In general, bioinformatics tools for integrative analyses can be classified as statistical methods, machine learning approaches and the hybrid of the two, including Bayesian approaches, probabilistic mixture models, maximum likelihood estimation, probabilistic graphical models, diffusion models, network analysis, etc. Various computational tools have been developed for integrative data analysis pipelines from variant detection to annotation and interpretation [8]. [9] discusses approaches that effectively weight and integrate hundreds of heterogeneous datasets into a regularized Bayesian integration system, and provides maps of function activities and interaction networks in more than 200 areas of human cellular biology.

Biomedical research has become a data intensive field, which enables data-driven approaches and requires sophisticated data mining and data and knowledge integration methods. However, most current data-driven and integrative approaches are majorly based on statistical metrics for the evaluations of data mining models. Due to lack of rigorous validations and systematic interpretation, there are a number of incidental findings and inconsistencies [10] in this field. What's more, it's difficult to rigorously evaluate numerous computational tools abovementioned.

Data and Knowledge Engineering

With the "omics" data explosion, biomedical sciences become data-intensive field, where data-driven approaches together with knowledge-driven approaches show great potential in knowledge engineering. While knowledge-driven approaches generate hypotheses from domain knowledge, data-driven approaches generate hypotheses by using computational methods with inductive learning.

Knowledge engineering deals with knowledge representation, automatic reasoning, statistical and mathematical methods, cognitive science and so on, and can contribute to clinical decision making. [16] provides a good review of knowledge engineering in translational bioinformatics. Biomedical data and knowledge sources have multiple formats, include numerical values, images, text, data streams, event logs, etc. Text mining, temporal data mining, workflow mining, network approaches have been and are still been extensively studied to mine these data sources. Databases, annotation systems, and biomedical ontologies should be combined together. Intelligent data analysis [3] has provided a way. Great efforts have been made to integrate domain knowledge and data sources using ontology and knowledge repositories.

Recent years have witnessed great achievements in technological engineering, such as electrical engineering and software engineering. However, knowledge engineering still lacks breakthrough. In the future, biomedical and translational sciences could have great potential to drive the development of knowledge engineering and vice versa. In the next section we will also discuss the significant potential of using reverse engineering to study complex biological systems.

INTERDISCIPLINARY CHALLENGES AND FUTURE DIRECTIONS

Translational bioinformatics deals with biology, medicine, mathematics, bioinformatics, clinical science, and so on, and is inherently interdisciplinary. There are great challenges as well as opportunities in multiple areas to push TBI forward.

Mathematical Foundations and Information Theory

Though advanced biotechnology enables us to accumulate terabytes and even petabytes of biomedical data, we cannot neglect the fact that the current accumulated data and data analysis is not sufficient to understand the complex biological systems. What's more, the mathematical foundation should be further enhanced for information extraction and knowledge discovery in biomedical research.

In general, the challenge comes with the distributed nature of data and knowledge sources [16]. Though we try to make a huge repository to store and manage all the data, we may never be able to store the data in a single repository with the dramatic increase of data every day. Another issue deals with data quality. Even though our technology is much better than before, we cannot guarantee that the measurements we made are accurate enough, not to say infinitely precise. The noises in our measurements make it hard to extract true information.

For a specific topic, we may need to use a corresponding dataset. However, we are even not sure which data should be used. Even worse, we don't know what kind of data and measurements we should collect. Even though we know certain variables are crucial, we might not be able to measure them correctly due to lack of standards, environmental factors and technological limitations.

Here comes a big question: **To what extent can we extract information and discover knowledge from incomplete and inaccurate measurements?** For example, each human genome has three billion base pairs nucleotides, yet our datasets at most consist of tens of thousands of samples (patients and controls). Since the genomic sequence data is highly "redundant", how can we decipher the information from the highly noisy big data but only a few samples? Great efforts have been made in mathematics [4; 5]. However, these results are built upon some restrictive constraints and far from solving the challenges in biomedical research.

Systems Biology and Network Science

A key goal of biomedical research is to elucidate the complex network of gene interactions underlying complex traits such as common human diseases. Systems biology approaches which integrate multiple information sources in a systematic way are gaining great popularity. For example, [18] introduced a multistep procedure for identifying potential key drivers of complex traits that integrates DNA-variation and gene-expression data with other complex trait data. Ordering gene expression traits relative to one another and relative to other complex traits is achieved by systematically testing whether variations in DNA that lead to variations in relative transcript abundances statistically support an independent, causative or reactive function relative to the complex traits under consideration [18].

Still our understanding of common human diseases and how best to treat them is hampered by the complexity of the human system in which they are manifested. Unlike simple Mendelian disorders, common human diseases often originate from a more complex interplay between constellations of changes in DNA and a broad range of factors such as diet, age and exposure to environmental toxins [17]. For instance, [17] proposed to link molecular states to physiological ones through the reverse engineering of molecular networks that sense DNA and environmental perturbations and drive variations in physiological states associated with diseases.

To understand biology at the system level, we must examine the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism [12]. Properties of systems, such as robustness, emerge as central issues, and understanding these properties may have an impact on the future of medicine. Systems biology and network science have great potential to play a key role in understanding complex biological systems and boosting translational science.

Reverse Engineering and Knowledge Engineering

Advanced technologies and biology have extremely different physical implementations, but they are far more alike in system-level organization than is widely appreciated [6]. In fact the principles and techniques widely used in engineering can be applied to computational systems biology [11]. Both biology and engineering are driven by demand for robustness to uncertain environments, and often deal with noisy measurements and incomplete information. To address these issues, modular architectures are ubiquitous in both domains that are composed of elaborate hierarchies of protocols and layers of feedback regulation and thus can significantly improve the system robustness. Robustness, modularity, feedback, and fragility are common characteristics of both engineering and biological complexities [13].

Brute-force computational approaches are hopeless for complex systems involving protocols and feedback. The success of systems biology will certainly require modeling and simulation tools from engineering, where great achievements have been made to address the challenges of uncertainties and robustness. Researchers in robust control theory, dynamic systems, and related areas have been vigorously pursuing mathematics and software tools to address these issues, which could also apply to complex biological systems [6], though scaling to deal with large biological networks will still be a major challenge.

Another big idea to boost biomedical research is knowledge engineering. As discussed in Section 3, knowledge engineering has great potential to integrate data and knowledge sources in an automatic and systematic way.

CONCLUSION

Translational bioinformatics represents a natural framework to properly and effectively apply data mining and machine learning techniques across molecular and clinical realms in the clinical decision-making context. The increasing share of data and methods in TBI has brought great opportunities as well as new tough challenges. One of the major challenges facing TBI community is to build systematic integrative analysis framework that can take advantage of multi-dimensional data and incorporate various information and knowledge sources. Data-driven and knowledge-driven models have been proposed to integrate multiple "omics" data as well as biomedical knowledge sources. However, rigorous validation methodologies are still necessary to evaluate these computational models in both scientific and clinical contexts. System biology and knowledge engineering could have great potential to push TBI forward and ultimately boost health care and improve people's life quality. As we are moving towards an era in which the amount of data produced every year is increas-ing exponentially, the TBI community needs to embrace this complexity and find new methods of analyzing data, extracting information and discovering knowledge.

REFERENCES

[1] ALTMAN, R.B., 2012. Introduction to translational bioinformatics collection. PLoS Comput Biol 8, 12, e1002796. DOI= http://dx.doi.org/10.1371/journal.pcbi.1002796.

[2] BELLAZZI, R., DIOMIDOUS, M., SARKAR, I.N., TAKABAYASHI, K., ZIEGLER, A., and MCCRAY, A.T., 2011. Data Analysis and Data Mining: Current Issues in Biomedical Informatics. Methods of Information in Medicine 50, 6 (2011), 536-544. DOI= http://dx.doi.org/10.3414/me11-06-0002.

[3] BELLAZZI, R. and ZUPAN, B., 2008. Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics 77, 2 (Feb), 81-97. DOI= http://dx.doi.org/10.1016/j.ijmedinf.2006.11.006.

[4] CANDES, E.J., ROMBERG, J., and TAO, T., 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. Ieee Transactions on Information Theory 52, 2 (Feb), 489-509. DOI= http://dx.doi.org/Doi 10.1109/Tit.2005.862083.

[5] CANDES, E.J., ROMBERG, J.K., and TAO, T., 2006. Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics 59, 8 (Aug), 1207-1223. DOI= http://dx.doi.org/Doi 10.1002/Cpa.20124.

[6] CSETE, M.E. and DOYLE, J.C., 2002. Reverse engineering of biological complexity. Science 295, 5560 (Mar 1), 1664-1669. DOI= http://dx.doi.org/DOI 10.1126/science.1069981.

[7] DENNY, J.C., 2012. Chapter 13: Mining electronic health records in the genomics era. PLoS Comput Biol 8, 12, e1002823. DOI= http://dx.doi.org/10.1371/journal.pcbi.1002823.

[8] DING, L., WENDL, M.C., MCMICHAEL, J.F., and RAPHAEL, B.J., 2014. Expanding the computational toolbox for mining cancer genomes. Nat Rev Genet 15, 8 (08//print), 556-570. DOI= http://dx.doi.org/10.1038/nrg3767.

[9] GREENE, C.S. and TROYANSKAYA, O.G., 2012. Chapter 2: Data-driven view of disease biology. PLoS Comput Biol 8, 12, e1002816. DOI= http://dx.doi.org/10.1371/journal.pcbi.1002816.

[10] HAIBE-KAINS, B., EL-HACHEM, N., BIRKBAK, N.J., JIN, A.C., BECK, A.H., AERTS, H.J.W.L., and QUACKENBUSH, J., 2013. Inconsistency in large pharmacogenomic studies. Nature 504, 7480 (Dec 19), 389-+. DOI= http://dx.doi.org/Doi 10.1038/Nature12831.

[11] KITANO, H., 2002. Computational systems biology. Nature 420, 6912 (Nov), 206-210. DOI= http://dx.doi.org/10.1038/nature01254.

[12] KITANO, H., 2002. Systems biology: A brief overview. Science 295, 5560 (Mar 1), 1662-1664. DOI= http://dx.doi.org/DOI 10.1126/science.1069492.

[13] KITANO, H., 2004. Biological robustness. Nature Reviews Genetics 5, 11 (Nov), 826-837. DOI= http://dx.doi.org/10.1038/nrg1471.

[14] KRISTENSEN, V.N., LINGJOERDE, O.C., RUSSNES, H.G., VOLLAN, H.K.M., FRIGESSI, A., and BORRESEN-DALE, A.-L., 2014. Principles and methods of integrative genomic analyses in cancer. Nature Reviews Cancer 14, 5 (May), 299-313. DOI= http://dx.doi.org/10.1038/nrc3721.

[15] NOY, N.F., SHAH, N.H., WHETZEL, P.L., DAI, B., DORF, M., GRIFFITH, N., JONQUET, C., RUBIN, D.L., STOREY, M.A., CHUTE, C.G., and MUSEN, M.A., 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 37, Web Server issue (Jul), W170-173. DOI= http://dx.doi.org/10.1093/nar/gkp440.

[16] PAYNE, P.R., 2012. Chapter 1: Biomedical knowledge integration. PLoS Comput Biol 8, 12, e1002826. DOI= http://dx.doi.org/10.1371/journal.pcbi.1002826.

[17] SCHADT, E.E., 2009. Molecular networks as sensors and drivers of common human diseases. Nature 461, 7261 (Sep 10), 218-223. DOI= http://dx.doi.org/10.1038/nature08454.

[18] SCHADT, E.E., LAMB, J., YANG, X., ZHU, J., EDWARDS, S., GUHATHAKURTA, D., SIEBERTS, S.K., MONKS, S., REITMAN, M., ZHANG, C.S., LUM, P.Y., LEONARDSON, A., THIERINGER, R., METZGER, J.M., YANG, L.M., CASTLE, J., ZHU, H.Y., KASH, S.F., DRAKE, T.A., SACHS, A., and LUSIS, A.J., 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics 37, 7 (Jul), 710-717. DOI= http://dx.doi.org/10.1038/ng1589.

[19] SMITH, B., ASHBURNER, M., ROSSE, C., BARD, J., BUG, W., CEUSTERS, W., GOLDBERG, L.J., EILBECK, K., IRELAND, A., MUNGALL, C.J., LEONTIS, N., ROCCA-SERRA, P., RUTTENBERG, A., SANSONE, S.-A., SCHEUERMANN, R.H., SHAH, N., WHETZEL, P.L., LEWIS, S., and CONSORTIUM, O.B.I.,

SMITH, B., ASHBURNER, M., ROSSE, C., BARD, J., BUG, W., CEUSTERS, W., GOLDBERG, L.J., EILBECK, K., IRELAND, A., MUNGALL, C.J., LEONTIS, N., ROCCA-SERRA, P., RUTTENBERG, A., SANSONE, S.-A., SCHEUERMANN, R.H., SHAH, N., WHETZEL, P.L., LEWIS, S., and CONSORTIUM, O.B.I., 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25, 11 (Nov), 1251-1255. DOI= http://dx.doi.org/10.1038/nbt1346.

[20] STACEY, M. and MCGREGOR, C., 2007. Temporal abstraction in intelligent clinical data analysis: A survey. Artificial Intelligence in Medicine 39, 1 (Jan), 1-24. DOI= http://dx.doi.org/10.1016/j.artmed.2006.08.002.

[21] WANG, L., ZHANG, A., and RAMANATHAN, M., 2005. BioStar models of clinical and genomic data for biomedical data warehouse design. Int J Bioinform Res Appl 1, 1, 63-80.

[22] WHETZEL, P.L., NOY, N.F., SHAH, N.H., ALEXANDER, P.R., NYULAS, C., TUDORACHE, T., and MUSEN, M.A., 2011. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 39, Web Server issue (Jul), W541-545. DOI= http://dx.doi.org/10.1093/nar/gkr469.

Computer Scientist in Profile: **Yang Zhang**

Contributor: Amarda Shehu[1,2] [amarda@gmu.edu]
Dept. of [1]Computer Science, [2]Bioengineering
George Mason University, Fairfax, VA 22030
Tel: 703-993-4135 Fax: 703-993-1710

"Yes, I did have a band in college, and I still love music," says Yang, who now has his own highly-successful computational biology laboratory at University of Michigan, Ann Arbor. While most of us are quick to associate his name with top protein structure prediction servers, such as I-TASSER and QUARK, not many of us know that Yang's first passion was music. "When I was young," he recalls, "my dream was to roam around the world with my guitar." His dream did not last long, as he made a conscious decision to pursue a graduate degree after completing his undergraduate studies in physics in China. His Ph.D. thesis was on the interaction of elementary particles, such as quarks (that explains a lot). He was awarded the prestigious Humboldt fellowship, which allowed him to study quarks at the Free University in Berlin for two years.

Yang's research interests took a sudden, unplanned shift to biology when he read an article by Zhongcan Ouyang on the shape of membrane vesicles in 1999. He recalls being fascinated by the fact that the predicted results on the shape of membrane vescicles could be directly confirmed and viewed in the wet laboratory. That was something he found he had been missing in his previous studies. No one has ever seen quarks, and their existence can only be indirectly validated through high-energy particle collisions. The need to connect computation with experiment ultimately drove Yang to Ouyangs lab at the Chinese Academy of Science, where he spent the two next years on studying the elasticity of RNA and DNA molecules.

In 2001, Yang joined Jeffrey Skolnick's laboratory in the University at Buffalo via a recommendation by Ulrich Hansmann. It was in Skolnick's lab where Yang started to learn how to fold proteins in silico. He recalls his time in the lab as the "Golden Age" of his research, as he could concentrate fully on science without worrying on having to secure funding or preparing lectures and other teaching materials. In Skolnick's lab, Yang developed a number of computer algorithms, including TM-score, TM-align, and SPICKER, which are still widely used in the community for comparing and analyzing protein structures. The most recognized accomplishment of his time in the Skolnick's lab is perhaps his TASSER method, which in essence allows assembling new protein structures from segments cut from known structures of other proteins. Using TASSER, Yang built the first genome-wide structure database of G protein-coupled receptors (GPCRs) in the human. This was a marked accomplishment, as GPCRs are now widely considered to be the most important and prevalent drug targets.

Yang continued his work on protein structure prediction and folding in Skolnick's lab till 2005. He then moved to the University of Kansas as an assistant professor. He and his team in Kansas continued Yang's journey on protein structure prediction and folding, as Yang fully recognized that his work was not done. "After more than forty years of effort," Yang says, "we still have not solved the problem of protein folding." In Kansas, Yang extended and improved TASSER to I-TASSER by iterative structure assembly simulations. He shared I-TASSER with the community through a web interface, and this resulted in I-TASSER establishing itself as one of the most widely used online structure prediction services. Since its development, I-TASSER has been consistently ranked as the best server for structure prediction in the community-wide "Critical Assessment of protein Structure Prediction" (CASP) experiment since 2006. The server has attracted so far more than $50,000$ registered users, with hundreds of jobs waiting on the queue on any single day. I recall having sent many of my undergraduate and graduate students over the years to Yang's I-TASSER server to complete their homeworks and their understanding of protein structure prediction. The server capabilities combine an intuitive and easy-to-use interface with serious algorithmic power and rigorous analysis.

"I-TASSER starts with a technique called threading, which requires the availability of homologous proteins," explains Yang. After I-TASSER, Yang wanted to do more and move to the *ab-initio* folding territory. To build a protein structure from scratch, Yang and his colleague, Dong Xu, developed a new algorithm based on "continuous fragment assembly,"' which Yang named QUARK. There seems to be a deeper story behind the naming than the subject of Yang's Ph.D. thesis. Yang explains, "In particle physics, hadrons, such as protons and neutrons, which account for the majority of the mass of all materials, are an assembly of quarks." In Yang's view, all protein molecules are an ordered reassembly of atomic building blocks (backbone fragments and side-chains), which is exactly the principle that QUARK follows in assembling structure models of novel protein sequences. QUARK made a debut worthy of its name. As soon as it was introduced to the community, QUARK stood out as the top *ab-initio* folding algorithm in the 9th and 10th CASP experiments.

In 2009, Yang moved his lab to the University of Michigan in Ann Arbor and joined the Department of Computational Medicine and Bioinformatics founded by Gil Omenn and Brian Athey. "I love what I am doing here and enjoy strong support from the department and colleagues," he says. He notes that one of the benefits of his new work environment is the ability to always find computational and experimental collaborators drawn to the same scientific quests as him. That has helped him pursue many projects and expand them from the silica to the wet laboratory.

"I am excited in particular by two major puzzles that we are now trying to solve in my lab: (1) What we can tell on a proteins role in cell when we are given the protein structure (mostly by computational prediction)? (2) How can we do the reverse of protein folding, i.e., design new protein sequences when given target structures?" Yang got started on the first puzzle by developing COFACTOR and COACH with his colleagues Ambrish Roy and Jianyi Yang. The programs detect drug- and ligand-binding partners from predicted structure models of proteins. The algorithms currently ranked at the top in the community-wide protein function annotation experiments (including CASP and CAMEO). To address the second second puzzle, Yang and his colleague, David Shultis, built up a new wet laboratory to crystallize proteins Yang designed in silico. They enjoyed a recent success in redesigning the BIR3 domain of the functional X-linked inhibitor of apoptosis, and the Phox membrane scaffolding domain; with the latter recently deposited in the Protein Data Bank.

"I consider it my duty to better serve the scientific community. If I do not share my inventions, they are essentially useless."

Despite the rapid string of successes, Yang does not forget what he considers his basic duty as a scientist. He and his team maintain a comprehensive set of web-based services for a variety of projects, ranging from protein structure prediction, protein function annotation, protein-protein interaction, and protein-ligand docking and drug screening. "The maintenance of multiple high-quality service systems can be time-consuming but worthwhile", Yang adds, "as one of our major goals when designing new algorithms is ultimately to better serve the scientific community through them."

"Going back to my first passion," he says, "I still play guitar and drum. I play drum at least 20 minutes every day, which helps me refresh my brain for a while from the crowding world of protein folding." He also routinely listens to Jazz. "But playing guitar or listening to Jazz are more enjoyable at night or during weekends, when it is quieter," he adds. Nurturing his first passion seems to be working for Yang, and, on that note, we have come to a natural conclusion.

Yang on one of his drumming sessions.

HealthGIS 2014

November 4 **Dallas, Texas**

Third ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health

In conjuncture with the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems

Call for Papers

OBJECTIVES

This workshop will provide a forum for researchers and practitioners to share new ideas and techniques for health-related GIS applications. We invite submission of original research related to all aspects of GIS usage and applications in medical and in healthcare systems. We especially encourage papers based on real-world experience.

WEB SITE

http://healthgis.tamu.edu/

IMPORTANT DATES

Paper Submission:	**September 5, 2014**
Notification of Acceptance:	September 19, 2014
Camera Ready:	October 3, 2014
Workshop:	November 4, 2014

TOPICS

Topics of interest for this workshop include but are not limited to the following:

- GIS applications to assist people in need, including disabled and elderly people
- Statistical analysis of environmental or medical spatial information
- Visualization of statistical spatial health-related data
- Privacy and security in location-based health-related systems
- Geographically-based access control for medical systems
- Location-based social networks as a tool to detect the spread of infectious diseases
- Planning evacuation routes for cases of mass disasters or for daily transport of people who need emergent medical treatment
- Using mobile computing and GPS in healthcare
- Management of location-based healthcare services

PAPER SUBMISSION

Submitted papers must not substantially overlap with papers that have been published or that are simultaneously submitted to a journal or a conference with proceedings. Submitted papers can be of three types:

- *Regular Research Papers*: these papers should report original research results or significant case studies. They should be at most 10 pages.
- *Case Study Papers*: these papers should provide a thorough scientific analysis of cases that are of interest to the research community. They should be at most 6 pages.
- *Position Papers*: these papers should report novel research directions or identify challenging problems. They should be at most 4 pages.

Paper should be submitted via the workshop submission page in the following link:
https://www.easychair.org/conferences/?conf=healthgis2014
See instruction in the Web page:
http://healthgis.tamu.edu/Submission.aspx

Authors of accepted papers must guarantee that at least one of the authors will attend the workshop and present their paper. The workshop proceedings will be included in the ACM Digital Library.

ORGANIZERS

- Daniel W. Goldberg, Texas A&M University (USA)
- Ori Gudes, Curtin University (Australia)
- Yaron Kanza, Jacobs Technion-Cornell Innovation Institute, Cornell Tech (USA)

SIGBIO Record - Submission Guidelines

Submission categories

Submissions to the newsletter can be either on a special issue topic or on topics of general interest to the SIGBIO community.

These can be in any one of the following categories:

- Survey/tutorial articles (short) on important topics.
- Topical articles on problems and challenges
- Well-articulated position papers.
- Review articles of technical books, products and .
- Reviews/summaries from conferences, panels and special meetings within 1 to 4 pages [1500-2500 words]
- Book reviews and reports on relevant published technical books
- PhD dissertation abstracts not exceeding 10 pages
- Calls and announcements for conferences and journals not exceeding 1 page
- News items on the order of 1-3 paragraphs

Brief announcements Announcements not exceeding 5 lines in length can simply be sent as ASCII text to the editors by e-mail. SIGBIO Record publishes announcements that are submitted as is without review.

Announcements cannot be advertisements and should be of general interest to the wider community. The Editor reserves the right to reject any requests for announcements at his discretion.

Authors are invited to submit original research papers or review papers in all areas of bioinformatics and computational biology. The papers published in SIGBioinformatics Record will be archived in ACM Digital Library. Papers should follow the ACM format, and there is no page limitation.

http://www.acm.org/sigs/publications/proceedings-templates

Submissions should be made via email to the editors Pierangelo Veltri *(University Magna Graecia of Catanzaro, Italy),*
Young-Rae Cho *(Baylor University)*