Translational Systems Genomics: Ontology and Imaging

Su-Shing Chen [1,,2,4] and Yu-Ping Wang [3]
1University of Florida-Gainesville, USA
2Partner Institute of Computational Biology,
Shanghai Institutes of Biological Sciences,
Chinese Academy of Sciences, China
3University of Missouri-Kansas City, USA
4corresponding author: suchen@picb.ac.cn

ABSTRACT

In developing an integrated framework for translational bioinformatics, we consider bioimaging in the NIH Roadmap that exploits high-resolution genomic imaging for clinical applications to the diagnosis and treatment of genetic disorders/diseases. On one hand, we develop new image processing techniques, while on the other, we use the fusion of several well known ontological standards - Gene Ontology (GO), Clinical Bioinformatics Ontology (CBO), Foundational Model of Anatomy (FMA) and Microarry Gene Expression Data Ontology (MGED) in this framework. We have discovered that the heterogeneity of the imaging data can be resolved at the different ontological levels of this framework. Moreover, structural genomic information can be readily integrated into the usual textual clinical information bases.

1. INTRODUCTION

A few years ago, when microarray imaging was first introduced, it was hailed to be "an array of hope" by Eric Lander, in Nature, 1999. But recently, it was considered as "an array of problems" by Frantz in Nature Review Drug Discovery, 2005. Currently, a considerable amount of research in genomics has focused on microarray gene expression analysis but little is being converted into clinical practice; it is well recognized that this functional imaging is strongly limited by poor reproducibility and accuracy. At the same time, high-resolution genetic probes evolved from the Human Genome Sequencing Project have been developed. When combined with imaging techniques, they provide high-resolution structural information about genomic variations [5-9]. These structural imaging techniques add an important extra dimension to the understanding of cell behavior and functioning for early disease diagnosis and drug response. However, these two independent sources of information, namely gene expression analysis and structural imaging, have never been correlated and used to enhance gene expression analysis. Therefore, our research group initiates to develop innovative computational imaging and statistical tools, which are capable of extracting and integrating structural/functional information, which will further expand the translational potential of the microarray technology in molecular diagnosis and personalized medicine [5-7].

2. APPROCHES

Our approach fills the gap between the development of high-resolution probes in genetics/genomics and the application of sophisticated computational imaging and statistical techniques. The research falls in line with the NIH Roadmap on Molecular Libraries and Imaging initiative[1] in that we intend to transform the emerging molecular imaging probes into clinical practice. After completion of the Human Genome Sequencing Project, the detection of genomic variations became a pressing issue, which was ranked as the No.1 scientific challenge in 2007 by Science Magazine. Specifically, the chromosomal anomalies are detected by the high-resolution structural imaging, while the resulting variations in the molecular function of genes are exploited by gene expression analysis. The structural and functional imaging information are complementary and at different resolution levels. Their combination will offer a comprehensive approach to characterize the complex traits (phenotypic information) of an organism because these phenotypic differences cannot be dictated by either structural or functional variations alone. However, implementing our hypothesis demands

[1] http://nihroadmap.nih.gov/molecularlibraries/.

a new paradigm to integrate multi-scale imaging information, i.e., to integrate molecular imaging with gene expression and to correlate them with phenotypic data. Furthermore, data from chromosomal structural abnormalities and gene expression are heterogeneous, making information integration or fusion from different imaging modalities computationally challenging. We turn to the recently developed clinical and gene ontology to remedy this heterogeneity. We gather the measurement of genetic and structural signals from three imaging modalities, e.g., gene expression microarray, aCGH probes and fluorescence in situ hybridization (FISH) (see Figure 2) by developing new image processing algorithms. The clinical applications will be the study of genetic disorders/diseases. Imaging based chromosome karyotyping has long been used as a more reliable tool than gene expression analysis. The former is currently being used in clinical cytogenetics laboratories. On the other hand, gene expression analysis provides complementary functional information. If combined, they offer a more accurate diagnosis methodology. We provide a "Systems Biology" approach to elucidate the complex traits of cancer and genetic disorders/diseases with structural/functional imaging at multiple resolutions. We believe that this is the first computational and quantitative approach to integrate the complementary aCGH, FISH and gene expression imaging information (Figures 1 and 5). We anticipate that the integrated and systematic approach will result in significant improvement over the current clinical genetic diagnostic procedures.

3. ONTOLOGICAL FRAMEWORKS AND PROTÉGÉ EXTENSIONS

In developing an integrated framework for biomedical informatics, the NIH Roadmap calls for the clinical and translational science knowledge management which requires the fusion of several well known ontological standards - Gene Ontology[2], Clinical Bioinformatics Ontology[3], Foundational Model of Anatomy[4], Microarray Gene Expression Data Ontology[5]. In this project, we explore how different ontologies can be integrated into a coherent knowledge management system for data mining from various heterogeneous image sources. With the exponential growth of biomedical data, biomedical researchers have met significantly a new challenge - how to exploit systematically the relationships between clinical and translational science data (e.g., genes and sequences) and the biomedical literature. Usually most of known genes are found in the biomedical literature and PUBMED is an important database for this kind of information. PUBMED, developed by the U.S. National Library of Medicine (NLM), is a database of indexed bibliographic citations and abstracts. It contains over 4,600 biomedical journals. PUBMED citations and abstracts are searchable via PUBMED[6] or the NLM Gateway[7]. The biomedical literature has much to say about gene sequence, but it also seems that sequence can tell us much about the biomedical literature. Currently, highly trained biologists read the literature and manually select appropriate Gene Ontology (GO) terms to annotate the literature with GO terms. Gene Ontology database has more recently been created to provide an ontological graph structure for biological process, cellular component, and molecular function of genomic data. McCray et al. have shown that the GO is suitable as a resource for natural language processing (NLP) applications because a large percentage (79%) of the GO terms has passed the NLP parser [2]. They also show that 35% of the GO terms were found in a corpus collected from the PUBMED database and 27% of the GO terms were found in the current edition of the Unified Medical Language System (UMLS). We have started to investigate them so that image data mining can be performed systematically in these domains. We analyze the Gene Ontology (GO), the Clinical Bioinformatics Ontology (CBO) and the Microarray Ontology (MO), explore the intersection of these three domains, and try to reason about the new information gained by combining them in Protégé using an in-house PHP/MYSQL tool that implements the inferencing and reasoning module. Ontology describes the basic categories and relationships of image data. In addition to this it defines entities and types of entities within its framework. It usually includes a vocabulary of terms where there are names for concepts,

[2] http://www.geneontology.org
[3] https://www.clinbioinformatics.org/cbopublic/
[4] http://sig.biostr.washington.edu/projects/fm/AboutFM.html
[5] http://mged.sourceforge.net/ontologies/index.php
[6] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
[7] http://gateway.nlm.nih.gov/gw/Cmd

definitions and defined logical relationships to each other. There is an important ontology system, Protégé[8], which is a free, open source ontology editor and knowledge-base framework developed by the NIH National Center for Computational Biology at Stanford. The Protégé platform supports two main ways of modeling ontologies via the Protégé-Frames and Protégé-OWL editors. Protégé ontologies can be exported into a variety of formats including RDF(S), OWL, and XML Schema. The Gene Ontology, Clinical Bioinformatics Ontology and Microarray Ontology are structured as directed acyclic graphs (DAGs). The terms can have one or more parents and zero, one or more children. Terms are linked by the is-a and part-of relationships.

Importing the three ontologies into a unified framework was essential to the process of information integration. Biomedical ontologies are being developed in ever growing numbers. Unfortunately there is still too little attention paid by the various separate groups involved to results already obtained by other groups working in neighboring or even overlapping fields. Therefore importing these three ontologies into a single framework, Protégé, was an attempt to start ameliorating this problem. The first module of this project involved importing all three ontologies into Protégé via OWL files, and the secondary modules for inferencing or reasoning are essential to integrate image data at the heterogeneous levels. Some simple samples are depicted in figures 3 and 4. The details will be published elsewhere.

4. CONCLUSIONS

This paper develops an integrated framework for translational bioinformatics in the area of bioimaging that exploits high-resolution genomic imaging for the clinical applications to the diagnosis and treatment of genetic disorders/diseases, including cancers. On one hand, we develop new image processing techniques, while on the other, we use the fusion of several well known ontological standards - Gene Ontology (GO), Clinical Bioinformatics Ontology (CBO), Foundational Model of Anatomy (FMA) and Microarry Gene Expression Data (MGED) Ontology in this framework. Interestingly, the heterogeneity of the imaging can be resolved at the different ontological levels of the framework. Moreover, structural genomic information can be readily integrated into the usual textual clinical information bases [2].

[8] http://protege.stanford.edu/.

5. REFERENCES

1.Yandell, M.D. and Majoros, W.H. (2002) Genomics and natural language processing. In Nature Reviews Genetics, 3(8), pp. 601-610.

2. McCray A.T., Browne A.C., and Bodenreider O. (2002) The Lexical Properties of the Gene Ontology (GO). In Proc. of AMIA Annual Symposium, pp 504-508.

3. Raychaudhuri, S., Chang, J. T., Sutphin, P.D., and Altman R.B. (2002) Associating Genes with Gene Ontology Codes Using a Maximum Entropy Analysis of Biomedical Literature. 12(1), January, pp 203-214.

4. Smith, B., Williams J., and Schulze-Kremer, S. (2003) The Ontology of the Gene Ontology, In Biomedical and Health Informatics: From Foundations to Applications, In Proceedings of the Annual Symposium of the American Medical Informatics Association, Washington DC.

5. Y.-P.Wang, M. Gunampally, J. Chen, D. Bittel, M. Butler and W.-W. Cai, A Comparison of Fuzzy Clustering Approaches for Quantification of Microarray Gene Expression, *Journal of VLSI Signal Processing Special Issue on Machine Learning for Microarray and Sequence Analysis*, 50: 305-320, 2008.

6. Yu-Ping Wang, Integration of GeneExpression and Gene Copy Number Variations with Independent Component Analysis,*30th Annual International IEEE EMBS Conference of the IEEE Engineering in Medicine and Biology Society in Vancouver, British Columbia, Canada during August 20-24, 2008.*

7. J. Chen and Yu-Ping Wang Detection of DNA copy number changes using statistical change point analysis, *Proceedings of the IEEE International Workshop on Genomic Signal Processing (GENSIPS), 2006*, May, College Station, TX.

8.Yu-PingWang and Ken Castleman, Automated Registration of Multi-Color Fluorescence In Situ Hybridization (M-FISH) Images for Improving Color Karyotyping, *Cytometry, Part A, 64A(2), April, 2005.*

9. Yu-Ping Wang and Wei-Wen Cai, Genetic imaging: where imaging science meets cytogenetic research, *Biophotonics Magazine, Nov., 2004*.

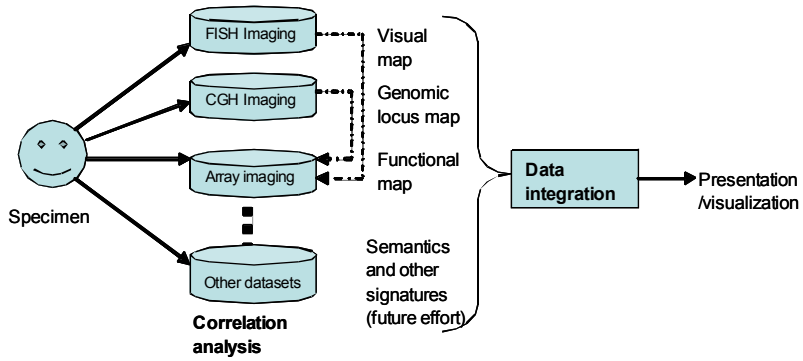

Figure 1 An overview of the proposed framework. Much of the ontological data is included in "other datasets" of semantic correlation analysis.

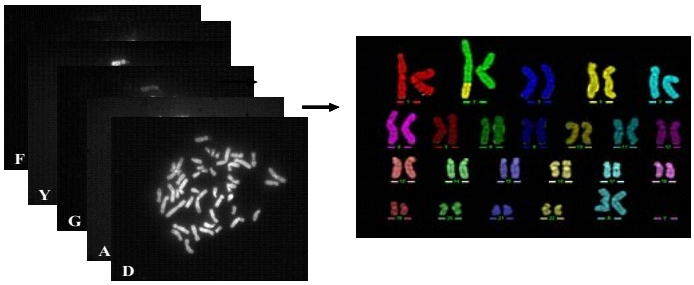

Figure 2 High resolution multi-color FISH probes are used for the detection of complex chromosomal abnormalities such as translocation (color karyotyping).

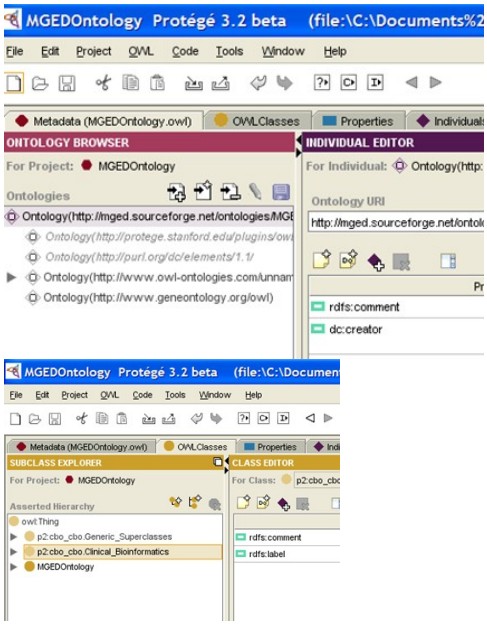

Figure 3 Metadata View and OWL Class View in the Protégé system.

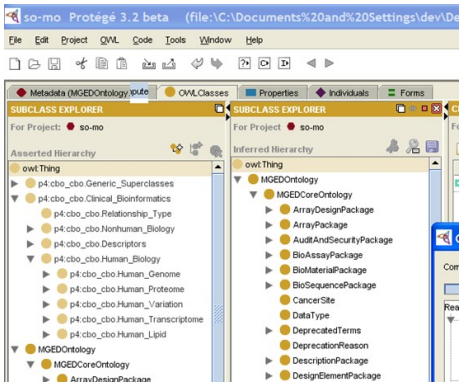

Figure 4 Inferred Hierarchy in the Protégé system, which will be basis for information fusion.

Systems genomics driven by multi-scale imaging

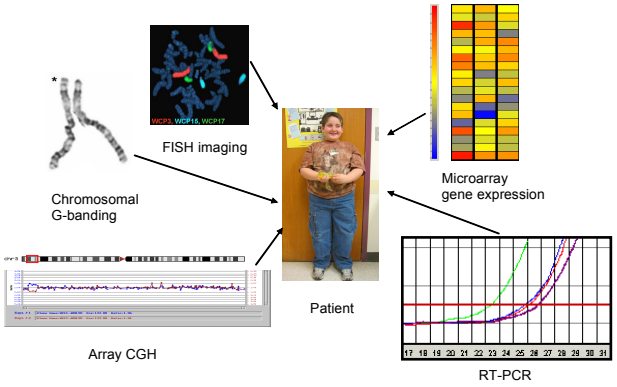

Figure 5. A comprehensive analysis of genomic variations from a patient with a chromosome 3p duplication using chromosome G banding, FISH imaging, BAC array CGH, quantitative RT-PCR and array gene expression approaches.