Accelerating Knowledge Discovery through Community Data Sharing and Integration

Y. L. Yip, Section Editor for the IMIA Yearbook Section on Bioinformatics
Swiss-Prot group, Swiss Institute of Bioinformatics, Geneva, Switzerland
University of Geneva, Dept. of Structural Biology and Bioinformatics, Geneva, Switzerland

Summary

Objectives: To summarize current excellent research in the field of bioinformatics.

Method: Synopsis of the articles selected for the IMIA Yearbook 2009.

Results: The selection process for this yearbook's section on Bioinformatics results in six excellent articles highlighting several important trends First, it can be noted that Semantic Web technology continues to play an important role in heterogeneous data integration. Novel applications also put more emphasis on its ability to make logical inferences leading to new insights and discoveries. Second, translational research, due to its complex nature, increasingly relies on collective intelligence made available through the adoption of community-defined protocols or software architectures for secure data annotation, sharing and analysis. Advances in systems biology, bio-ontologies and text-ming can also be noted.

Conclusions: Current biomedical research gradually evolves towards an environment characterized by intensive collaboration and more sophisticated knowledge processing activities. Enabling technologies, either Semantic Web or other solutions, are expected to play an increasingly important role in generating new knowledge in the foreseeable future.

Keywords

Medical informatics, International Medical Informatics Association, yearbook, bioinformatics, translational research, knowledge discovery

Yearb Med Inform 2009:117-20

Introduction

Efficient data sharing and integration is at the center of fruitful interdisciplinary and translational research. In the 2008 Yearbook, it was already noted that effort was made in the development of data standards or representation models to better manage the increasing amount of biomedical data [1] using either Semantic Web technology [2] or other methods [3]. Recent survey of the literature shows that these efforts and trends continue. Within the Semantic Web technology framework, it is fascinating to note that a range of solutions based on RDF (a standard format for document) and OWL (a language for ontology specification) have been implemented to reap the benefits promised by the Semantic Web [4-8]. Other community-defined and adopted protocols, software tools or infrastructures, such as caGRID, REDCap, DAS, are also gaining momentum throughout the year [9-13]. These software architectures enable secure web-based data annotation or sharing, and thus facilitate multi-institutional collaborative works and data analysis.

With the accrued ease to access data of all nature, privacy surrounding an individual's genetic data continues to be a major topic of discussion. In 2008, this discussion was further fueled by the discovery by Homer *et al.*, who demonstrated that there was a risk of disclosure of an individual's identity even with summary-level genetic data, such as statistics [14]. This finding, together with the advent of next-generation genome sequencing technique giving easy access to human genomes [15], guarantee that data protection, disclosure, ethic will be the ever pressing issues to be resolved both at the political level and at the informatics handling level.

In 2008, continual advancement in the application domains of systems biology [16-17], bio-ontologies [18-20], and text-mining [21-22] are also noted.

Best Paper Selection

The best paper selection of articles for the section 'bioinformatics' in the IMIA Yearbook 2009 follows the tradition of previous yearbooks [1, 23-24] in presenting examples of excellent research reflecting the above-mentioned trends. Six articles were selected this year as a result of a comprehensive review process. Three of these papers directly reflect the use of Semantic Web to realize heterogeneous data integration and knowledge inference [6-8]. Belleau *et al.* reported a tool to convert bioinformatics data and knowledgebases to RDF format [6]. Their tool facilitates light-weighted approaches that enable the linking of data resources using RDF triples, and accelerate the creation of data mashups. Splendiani, on the other hand, offered a plugin that extended a specialized software analysis platform (Cytoscape) with support for reasoning on ontologies in the semantic web framework [7]. All three papers illustrated the capabilities in RDF/OWL to further knowledge discovery through inference with concrete examples [6-8]. Two other papers in the best paper selection presented community-based data sharing solutions that enable more efficient multi-institutional collaborative works

[10,12]. The last paper by *Homer et al.,* as mentioned in Introduction, challenges directly our confidence in regard to our genetic privacy [14].

Table 1 presents the selected papers. A brief content summary of the selected best papers can be found in the appendix of this report.

It is worth noting that the best paper selection in the field of Bioinformatics is becoming ever more challenging due to the large number of articles and their presence in more specialized journals (e.g. neuroscience journals, human genetic journals). Indeed, in year 2008 alone, the number of articles having a mention of Bioinformatics in their medical subject heading reached about 8000. Therefore, although the selected papers give an overview of important trends and challenges in topics of Bioinformatics that are most relevant to medical informatics, they certainly do not cover all the aspects of this broad field.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2009 in the section 'Bioinformatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Bioinformatics
▪ Amin W, Parwani AV, Schmandt L, Mohanty SK, Farhat G, Pople AK, Winters SB, Whelan NB, Schneider AM, Milnes JT, Valdivieso FA, Feldman M, Pass HI, Dhir R, Melamed J, Becich MJ. National Mesothelioma Virtual Bank: a standard based biospecimen and clinical data resource to enhance translational research. BMC Cancer 2008 Aug 13;8:236.
▪ Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 2008 Oct;41(5):706-16.
▪ Gudivada RC, Qu XA, Chen J, Jegga AG, Neumann EK, Aronow BJ. Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge. J Biomed Inform 2008 Oct;41(5):717-29.
▪ Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)- a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-81.
▪ Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008 Aug 29;4(8):e1000167.
▪ Splendiani A. RDFScape: Semantic Web meets systems biology. BMC Bioinformatics 2008 Apr 25;9 Suppl 4:S6.

Conclusions and Outlook

The best paper selection for the Yearbook section 'Bioinformatics' shows that effort to accelerate translational research via more efficient data integration and analysis strategies had continued in 2008. Although there is still a widespread impression that the Semantic Web RDF/OWL standards have not attained maturity, our selection reveals that numerous applied solutions are being offered and the application of logic to infer new insights is gradually becoming a reachable reality. Researchers are also more accustomed to web-based, or virtual system for data annotation and sharing. There is a noticeable progression from web of services to web of people linked together by common research focus or interest. In this context, there will be a need to collaborate between communities and incorporate current existing standards. For example, it may be necessary for the Semantic Web community to reach out to the HL7 community, a reference of the clinical informatics area, so that one can further close the loops in the biomedical area.

Acknowledgement
I would like to acknowledge the support of Martina Hutter and the reviewers in the selection process of the IMIA Yearbook.

References

1. Yip YL. The promise of systems biology in clinical applications. Findings from the Yearbook 2008 Section in Bioinformatics. Yearb Med Inform 2008;102-4.
2. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with semantic web. BMC Bioinformatics 2007;8:S2.
3. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40: 30-43.
4. Baker CJO, Cheung KH, editors. Semantic Web: revolutionizing knowledge discovery in the life sciences. New York: Springer; 2007.
5. Deus HF, Stanislaus R, Veiga DF, Behrens C, Wistuba II, Minna JD, et al. A semantic web management model for integrative biomedical informatics. PLOS one 2008;3(8):e2946.
6. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 2008;41:706-16.
7. Splendiani A. RDFScape: Semandtic Web meets Systems Biology. BMC Bioinformatics 2008,9 (Suppl 4):S6.
8. Gudivada RC, Qu XA, Chen J, Jegga AG, Neumann EK, Aronow BJ. Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge. J Biomed Inform 2008;41:717-29.
9. Oster S, Langella S, Hastings S, Ervin D, Madduri R, Kurc T, et al. caGrid 1.0: a grid enterprise architecture for cancer research. AMIA Annual Symposium Proceedings 2007: 573-7.
10. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009; 42(2):377-81. Epub 2008 Sep 30.
11. Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, et al. Integrating biological data – the Distributed Annotation System. BMC Bioinformatics 2008;9 (Suppl 8):S3.
12. Amin W, Parwani AV, Schmandt L, Mohanty SK, Farhat G, Pople AK, et al. National Mesothelioma Virtual Bank: A standard based biospecimen and clinical data resource to enhance translational research. BMC Cancer 2008;8:236.
13. Sabb FW, Bearden CE, Glahn DC, Parker DS, Freimer N, Bilder RM. A collaborative knowledge base for cognitive phenomics. Mol Psychiatry 2008;13:350-60.
14. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individual contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotype microarrays. PLOS Genet 2008,4(8):e1000167.
15. Voelkerding KV, Dames SA, Durtschi JD. Next generation sequencing: from basic research to diagnostics. Clin Chem 2009;55(4):641-58.
16. Kim PS, Lee PP, Levy D. Dynamics and potential impact of the immune response to chronic myelogenous leukemia. PLOS Comput Biol 2008;4(6): e1000095.
17. Banaji M, Mallet A, Elwell CD, Nicholls P, Cooper CE. A model of brain circulation and metabolism: NIRS signal changes during physiological

challenges. PLOS computational Biology 2008; 4(11):e1000212.
18. Viti F, Merelli I, Caprera A, Lazzari B, Stella A, Milanesi L. Ontology-based, tissue microarray oriented, image centered tissue bank. BMC Bioinformatics 2008;9 (Suppl 4):S4.
19. Coulet A, Smäil-Tabbone M, Benlian P, Napoli A, Devignes MD. Ontology-guided data preparation for discovering genotype-phenotype relationships. BMC Bioinformatics 2008;9 (Suppl 4): S3.
20. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 2008;83(5):610-15.
21. Theodosiou T, Angelis L, Vakali A. Non-linear correlation of content and metadata information extracted from biomedical information extracted from biomedical article datasets. J Biomed Inform 2008;41:202-16.
22. Roberts A, Gaizaukas R, Hepple M, Guo Y. Mining clinical relationships from patient narratives. BMC Bioinformatics 2008;9 (Suppl 11):S3.
23. Lang E. Bioinformatics and its Impact on Clinical Research Methods. Methods Inf Med 2006;45:104-6
24. Lang E. Integrating bioinformatics into clinical practice: progress and evaluation. Methods Inf Med 2007;46:106-8.

Correspondence to:
Dr. Yum Lina Yip
Swiss-Prot group
Swiss Institute of Bioinformatics
1 Rue Michel-Servet
CH-1211 Geneva, Switzerland
Tel: +41 22 379 5049
Fax: +41 22 379 5858
E-mail: lina.yip@isb-sib.ch

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2009, Section Bioinformatics*

Amin W, Parwani AV, Schmandt L, Mohanty SK, Farhat G, Pople AK, Winters SB, Whelan NB, Schneider AM, Milnes JT, Valdivieso FA, Feldman M, Pass HI, Dhir R, Melamed J, Becich MJ

National Mesothelioma Virtual Bank: A standard based biospecimen and clinical data resource to enhance translational research

BMC Cancer 2008;8:236

The recent advancements in translational medicine have led to a need for biobanks with high quality and well-annotated tissue samples that can meet the demand of the basic, clinical and translational research communities. This paper presents the National Mesothelioma Virtual Bank (NMVB), a virtual biospecimen registry with robust informatics support that facilitates management, standardized collection and detailed clinical annotation of Mesothelioma cases across multiple collaborative sites. The architecture of the NMVB is based on three major components: (a) common data elements that provide semantic and syntactic interoperability across multiple institutions to facilitate translation research, (b) clinical and epidemiologic data annotation, and (c) data query tools. The data managers at each collaborative site are responsible for clinical annotation of biospecimen and collect related information from other data sources. The information is then integrated, de-identified and standardized. Patient privacy is protected as the web-based query tool only discloses de-identified information to the end users. The NMVB currently has over 600 annotated cases that include paraffin embedded tissues, tissue microarrays, blood and DNA samples.

Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J

Bio2RDF: Towards a mashup to build bioinformatics knowledge systems

J Biomed Inform 2008;41:706-16

Within the Semantic Web framework, RDF was proposed as a standard format together with ontologies encoding their semantics. This format is however not yet common in the web. This paper described an open source project, Bio2RDF, which offers a tool to convert bioinformatics data and knowledge bases to RDF format. As such, Bio2RDF can act as a mashup system to help bioinformatics knowledge integration by providing access to RDF documents from many different resources linked together with normalized URIs and a common ontology. The paper provided a description of the Bio2RDF three-step approach to build mashups. It also illustrated the use of a mashup to explore the implication of four transcription factor genes in Parkinson's disease. The knowledge space created by Bio2RDF is directly usable by a true semantic web browser such as Tabulator. Bio2RDF is also extensible and flexible in that it offers users the possibility to add knowledge sources or experimental private data by creating new rdfizers. Privacy of private data can be assured by using its built-in routing capability. The authors showed that it is possible to scale up to millions of documents. As a work in progress, Bio2RDF's ontology and its rdfizer programs are not definitive, and the authors welcome contribution from the community.

Gudivada RC, Qu XA, Chen J, Jegga AG, Neumann EK, Aronow BJ

Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge

J Biomed Inform 2008;41:717-29

The identification of genes responsible for causing or preventing human disease is essential for understanding the pathophysiological mechanisms and developing new diagnostics and therapeutics. In this paper, the authors described, for the first time, an approach to prioritize candidate genes using Semantic Web standards and techniques. In their method, W3C's RDF and OWL standards were first used to integrate genomic and phenomic annotations associated with the candidate gene set. Centrality analysis was then performed on the resulting network data structure (a directed acyclic graph) to rank genes according to the model-driven semantic relationships. The method was benchmarked against 60 diseases and was applied to prioritize candidate genes from cardiovascular diseases. It was shown to be able to uncover relations

* The complete papers can be accessed in the Yearbook's full electronic version, provided that permission has been granted by the copyright holder(s).

and centrality elements that could lead to specific hypotheses and new insights. The Semantic Web approach offers the flexibility to add more knowledge resource to enhance the method's overall performance. The current limitations are that the prioritization accuracy relies on the underlying online sources, and that it can be applied only on diseases where clinical features are available. The authors believed that their methods can be applied to all OMIM diseases having known loci but unknown molecular basis.

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG

Research electronic data capture (REDCap)- A metadata-driven methodology and workflow process for providing translational research informatics support

J Biomed Inform 2009,42(2):377-81

Biomedical research will increasingly involve collaborations by many scientists in diverse locations connected securely through computer networks that enable data submission, analysis, and sharing. To address this need, this paper reported a software toolset, Research electronic data capture (REDCap), which is designed to provide scientific research teams with an easy workflow to develop secure, web-based applications for collecting, storing and disseminating clinical and translational research data. The workflow to create a new REDCap project includes an initial meeting between researchers and REDCap informatics core representative for project-specific metadata definition and creation, a period for prototype testing and refinement, and then production. Based on this workflow, the researcher-led metadata creation process is flexible, the startup time to launch a new project is short and resource investment is little. REDCap provides: 1) an interface for data entry; 2) audit trails for tracking data manipulation and export procedures; 3) automated export procedures for data downloads; 4) procedures for importing data from external sources; and other advanced features. Despite certain limitations discussed in the paper, REDCap has gained collaborative support from a wide consortium of national and international partners. Indeed, from its initial development and deployment at Vanderbilt University (USA) in 2004, the number of translational research projects using REDCap and active partner institutions had increased to 286 and 27, respectively, by the time the paper was published (2008). Currently (May 2009), REDCap supports approximately 530 studies across an international consortium of 51 institutions.

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW

Resolving individual contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotype microarrays

PLOS Genet 2008,4(8):e1000167

It is a common perception that it is not possible to identify individuals using pooled data (e.g. allele frequency) from single nucleotide polymorphism (SNP) data. In this paper, the authors demonstrated an approach using high-density SNP genotyping microarray to accurately determine trace contributions (<0.1%) of an individual's genomic DNA to a complex DNA mixture. The authors first developed a theoretical framework for detecting an individual's presence within a mixture, used simulation to test the limitations of these approaches, and then experimentally demonstrated the identification of an individual's DNA within an highly-complexed assayed mixture. Described in simplistic terms, their method determines if a person is in a mixture by comparing two statistically describable distances: the distance of the individual from a reference population and the distance of an individual from the mixture. The implication of their finding in forensic applications and genome-wide association studies were discussed. In the latter case, there is a considerable effort to make experimental data publicly available so that the data can be pooled to provide summary-level statistics, in part to mask individual-level genotype data. With the current study, it is now clear that individuals can be identified based on summary-level statistics and further research is needed to determine how to best share data while respecting the privacy issues with genetic data.

Splendiani A

RDFScape: Semandtic Web meets Systems Biology

BMC Bioinformatics 2008;9 (Suppl 4): S6

In this paper, the authors presented RDFScape, a plugin that has been designed to extend a software platform for biological network analysis (Cytoscape) with support for reasoning on ontologies in the semantic web framework. RDFScape offers the possibility to query and visualize the information explicitly asserted in ontologies, and also the knowledge that can be inferred from them. The inference process can further be tuned by the users to produce their own interpretation of data. Two examples with ontologies relative to biological pathways were used to illustrate how these ontologies can be abstracted and visualized as interaction networks, and how inference can be used on pathways to answer specific queries. As such, RDFScape fills a gap in the availability of tools that rely on ontologies for biological data analysis. Although RDFScape presents limitations in the implementation of the semantic web which is proper to the early stage of new technology adoption, this paper demonstrates how ontologies and reasoners in the semantic web framework can already be used for real tasks.