Ontologies, semantic web and intelligent systems for genomics

Christopher J. O. Baker, Greg Butler, Volker Haarslev

Concordia University, Montreal, Canada

Abstract

'Ontologies, the semantic web and intelligent systems for genomics' is the first project of its kind in Canada to focus on bringing semantic web technology to genomics. Ontology, multi-agent systems, machine learning and natural language processing are used to build tailored knowledge base and semantic systems of direct use to the scientific discovery process. Major challenges of the post genomic era, namely data integration and knowledge retrieval are addressed.

Introduction

Since its inception bioinformatics has concerned itself with storage, management, and analysis of biologically relevant data derived from experimental and in-silico biological analysis. Initially such information was predominantly sequence information along with hand curated annotation. Now bioinformatics and genomics cover a wide range of data types that are stored, used, and manipulated by scientists and machines. Access to data is more complex, it is distributed in different locations, it is non-static, often being updated by newer versions of processed results. Users of such data are experts in a field of medical or biological study and are not necessarily technically skilled to access data relevant to their scientific needs. As a consequence further technologies that go beyond storage and processing of data are required to communicate the knowledge inherent in the data. Semantic access across the internet to data stored according to biological axioms is a growing need. The technological steps required to perform this are the considered goals of the Semantic Web community and there is a clear justification of developing a semantic web for genomic data. Some bioinformatics projects have identified the role of ontologies and semantics in their future successes. Notably the Gene Ontology Consortium [1] and TAMBIS [2], amongst others have sought to formally describe biological knowledge in the new OWL (Ontology Web Language) the w3c [3] adopted standard for formal ontology markup on the web. The Gene Ontology has promoted collaboration in genome annotation, and facilitated the exchange of data between communities working on different organisms. The TAMBIS database system demonstrated effective data integration by using formal ontologies, and demonstrated ease of querying a collection of data through the use of the ontology of terms familiar to scientists.

This paper outlines the Semantic Web research being carried out by the Fungal Web project 'Ontologies, the semantic web and intelligent systems for genomics' at Concordia and McGill Universities.

The semantic web for genomics:

The key technical requirements for the development of the semantic web for genomics include the provision of formal ontologies associated with web sites, automated agent systems, text mining technologies, and relational data analysis. Together these components can deliver a robust integrated platform to provide genomics knowledge through semantic access.

Semantic access is achieved through ontologies associated with web sites. Ontologies describe the information in a way that is precise and formal enough to be manipulated by reasoning software and query tools. Agents are integral to the semantic web: an agent is software that knows how to navigate the web and to use the available knowledge to achieve a given task. In order to achieve its task an agent may need to cooperate with other agents, it may need to learn more, and it may need to be proactive in response to changing data as the data becomes available on the web.
Text mining tools can retrieve highly specific information from free text for ontology instantiation to form a knowledge base. Without text mining, many valuable units of genomic annotation are only manually accessible within the rich contextual information of a scientific text and cannot contribute to a knowledge base. Relational data mining (RDM) approaches serve as the machine end user for the multiple units of distributed data made available by semantic web access and software agents.

RDM techniques such as Probabilistic Relational Modeling and Bayesian Networks have superior data mining and inference capabilities of use to the biologist.

Ontology

Biologists are becoming comfortable with using ontologies as extended annotation tools. As their utility is acknowledged the limitations of existing ontological domains are more pronounced. Building ontology of a new genomics domain is complex, particularly the knowledge acquisition step of the ontology development lifecycle. This step requires prolonged and frequent interaction with domain experts to define and formalize concepts and relationships that may not yet be described in text books. As bioinformatics data becomes more interdisciplinary, combining and integrating existing ontologies is a further focus of ontology development. The Fungal Web Ontology, written in OWL, covers enzyme classification, enzyme functional parameters, enzyme applications, and fungal taxonomy gathered from distributed resources. The Biologist's interest in ontologies goes beyond a declaration of concepts and relationships. The pre-structured knowledge of the ontology is valued for its query-answer utility, used in conjunction with simple query tools incorporated in ontology editor Protégé [4] or description logic reasoning tools such as Racer [5].

Agents

The increasing number of distributed, independent data sources increasingly inhibits scientists from conducting comprehensive bioinformatics analyses. We support the introduction of wrapper specific agents to wrap each data source to make these sources penetrable through multi agent systems [6]. We envision agents to cooperate and coordinate access to information from multiple sources to produce task specific data. This approach not only avoids building several monolithic applications but additionally allows the incorporation of further agent specific features like a proactive multi-process and update process whenever new results become available. Multi source retrieval permits new applications to exist such as the Mutation Miner [7] developed within the Fungal Web project. This application permits the annotation of protein structures with the free text annotations describing mutations made to protein sequence homologs. It depends upon coordinated access to free text, sequence and structure databases as well as distributed algorithmic tools.

Text Mining

Numerous resources required by scientists are provided only in free text format, particularly in the case of scientific publications. Whilst such documents are accessible in electronic formats, only specific information may be required from a text. Much time can be saved by the scientist if relevant papers are selected by automated information retrieval systems, parsed for specific information and databases automatically established and updated.

Tools for such analysis are coming of age and presently require tailoring for specific needs particularly in the biological domain. In light of the current trend where of funding for manual curation of databases is being reduced [8] text mining and natural language processing techniques have a golden opportunity to establish themselves at the heart of the bioinformatic annotation arena. To date these techniques have shown great promise in mining scientific literature most notably for named entity extraction e.g. gene and protein names, protein-protein interactions. Template, statistical and lexical based methods contribute to natural language processing and have been demonstrated to have different levels of resolution and granularity. Consequently one of the challenges involves matching the correct tool with the correct application and an assessment of the contribution each approach can provide.

Within the Fungal Web project there is considerable expertise with computational linguistic approaches that perform a range of sophisticated analyses of free text (tokenisation, gazetteering, named entity recognition, part of speech tagging, co-reference resolution, noun phrase chunking, co-location, predicate argument recognition). These approches are typically applied within the GATE framework (General Architecture for Text Analysis) [9] and a customized series of such analyses is prepared to suit the specific task. As these existing methods are being migrated to scientific text analysis reliability measures for such tools on a particular task are crucial to the successful implementation of natural language processing to bioinformatics. The Fungal Web group is currently developing automated methods to evaluate precision and recall for text mining analyses and characterizing issues unique to scientific texts and domains.

Relational Data Mining

In biology the trend towards relational data mining results from the need to consider multiple sources and types of evidence to model a biological phenomenon. Studies of gene regulation initially relied on gene mutation and knockout experiments before microarray technologies were developed and genomic information (transcription factor binging sites, promoter sequence) was available. Gene regulation occurs as a result of numerous interactions between proteins and DNA molecules resulting in the need to comprehend large causal networks. Probabilistic relational data mining techniques have recently been shown to provide superior insight to such networks permitting the reconstruction of regulatory networks and providing a rich source of probabilistic relations for validation in laboratory experiments [10].

Probabilistic relational modeling is a language based on relational logic for statistically describing complex domains in terms of multiple entities, their properties, and the dependency relations between them. In a PRM uncertain properties of an entity depend probabilistically both on other properties of that entity and on properties of related entities or the relational structure of the entities. Such models can be learned directly from existing databases using well-founded statistical techniques. Generic dependency models are built at the class level and instantiated for specific circumstances. Pattern recognition in uncertain data sets is possible with PMRs and unknown features can be inferred. Within the FungalWeb project PRMs are being employed to analyze fungal gene expression data sets.

Fungal Genomics

The Fungal Web project is affiliated with the Fungal Genomics project at Concordia [11] which is a large-scale, gene discovery program on evolutionarily diverse fungal species. This project plans to discover over 70,000 new fungal genes and develop high throughput methods to characterize the function of the enzymes produced by these genes. The effectiveness of these gene products in industrial processes and in environmental remediation is also tested. This data intensive project is supported by sophisticated bioinformatics resources with clearly defined needs. Integration of the tools developed within the Fungal Web project to the Fungal Genomic bioinformatics is a common goal of the two projects.

Conclusions

More than ever, the challenges of scientific research require the close collaboration of specialists from computational and scientific domains. Our project acknowledges this need and focuses on technologies to enhance the scientific discovery process. The project's successful conclusion will place Quebec at the vanguard, both within Canada and internationally, of intelligent systems for genomics, and in the exploitation of the semantic web for genomics.

Acknowledgements

Co-Principal investigators on the Fungal Web project include; Sabine Bergler, Leila Kosseim, Guy Lapalme, Doina Precup, Justin Powlowski, Adrian Tsang, and Nematollaah Shiri. Funding was generously provided by Genome Quebec.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. Nat Genet 2000 May; 25(1):25-9
2. Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R. TAMBIS--Transparent Access to Multiple Bioinformatics Information Sources. Proc Int Conf Intell Syst Mol Biol. 1998;6:25-34
3. Bechhofer S, van Harmelen F, Hendler J, Horrocks I, McGuinness DL, Patel-Schneider PF, Stein LA, OWL Web Ontology Language W3C Recommendation February 2004 http://www.w3.org/TR/2004/REC-owl-ref-20040210/
4. Noy N. F., Sintek M., Decker S., Crubezy M., Fergerson R. W., & Musen M. A.. Creating Semantic Web Contents with Protege-2000 IEEE Intelligent Systems 16(2):60-71, 2001.
5. Haarslev V, Möller R Description of the RACER System and its Applications. Proceedubgs International Workshop on Description Logics (DL-2001), Stanford, USA, 1.-3. August 2001.
6. Graham J, Windley V, McHugh D, McGeary F, Cleaver D, and Decker K, Tools for Developing and Monitoring Agents in Distributed Multi Agent Systems, Workshop on Agents in Industry at the Fourth International Conference on Autonomous Agents, Barcelona, Spain, June, 2000
7. Baker C.J.O. and Witte R. Enriching Protein Structure Visualizations with Mutation Annotations Obtained by Text Mining Protein Engineering Literature The 3rd Canadian Working Conference on Computational Biology (CCCB'04) Co-located with IBM CASCON conference, Markham, Ontario, October 2004
8. EMBOSS seeks funding as UK Human Genome Mapping Project Resource Centre closes down. `http://bioinformatics.org/forums/forum.php?forum_id=2663`, 2004.
9. Cunningham H. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36:223–254, 2002. `http://gate.ac.uk`.
10. Segal E, Rich Probabilistic Models for Genomic Data, Ph.D. Thesis, 2004 August, Stanford University
11. Fungal Genomics at Concordia https://fungalgenomics.concordia.ca/home/index.php