Editorial

Amandeep S. Sidhu, Tharam S. Dillon and Elizabeth Chang

Digital Ecosystems and Business Intelligence Institute,
Curtin University of Technology, Perth, Australia
E-mail: Amandeep.Sidhu@cbs.curtin.edu.au
E-mail: Tharam.Dillon@cbs.curtin.edu.au
E-mail: Elizabeth.Chang@cbs.curtin.edu.au

Jake Y. Chen

Indiana University School of Informatics, Indianapolis, IN
and
Department of Computer and Information Science,
Purdue University School of Science, Indianapolis, IN
E-mail: jakechen@iupui.edu

Biographical notes: Amandeep S. Sidhu is a senior researcher at Digital Ecosystems and Business Intelligence Institute at Curtin University of Technology, Perth, with expertise in Protein Informatics. He is currently leading an innovative Protein Ontology Project (since 2003). His research interests include biomedical ontologies, structural bioinformatics, proteomics, XML enabled web services, and artificial intelligence. His work in these fields resulted in over 30 scientific publications. He is currently involved with many semantic web and bioinformatics conferences and workshops as an organiser or as a programme committee member. He is current acting as a coordinator for the IEEE Bioinformatics and Systems Biology Community. He is also serving as Vice Chair of NSW Chapter of IEEE Engineering in Medicine and Biology Society.

Tharam S. Dillon is a Research Professor with Digital Ecosystems and Business Intelligence Institute at Curtin University of Technology, Perth. He is the Chair of Working Group on Web Semantics (WG 2.12/12.4) on TC2 for International Federation for Information Processing (IFIP), and Chair of IEEE/IES Technical Committee on Industrial Informatics. He is an expert in conceptual modelling, XML modelling, ontology development, data mining, and knowledge engineering. He has published six authored books and six co-edited books. He has also published over 600 scientific papers and has over 1000 citations of his work in refereed journals and conferences.

Elizabeth Chang is currently Director of the Digital Ecosystems and Business Intelligence Institute at Curtin Business School, Curtin University of Technology, in Perth, Western Australia. She is the Vice-Chair of the Work Group on Web Semantics (WG 2.12/12.4) in Technical Committee for Software: Theory and Practice (TC2) for the International Federation for Information Processing (IFIP). She has published over 300 scientific conference and journal papers including three authored books and two co-edited books. The themes of these papers are in the areas of ontology,

Copyright © 2007 Inderscience Enterprises Ltd.

software engineering, object/component based methodologies, e-commerce, trust management and security, web services, user interface and web engineering as well as logistics informatics.

Jake Y. Chen is an Assistant Professor of Informatics at Indiana University School of Informatics, and Assistant Professor of Computer Science at Purdue University School of Science Department of Computer and Information Science, Indianapolis, Indiana He is an IEEE senior member. He has been active in informatics R&D in the biotech industry and academia for the past ten years. He authored or co-authored more than 30 peer-reviewed articles in the area of bioinformatics, biological databases, and systems biology; and he made more than 50 invited presentations at international conferences, academic institutions, and companies worldwide. His industrial and entrepreneurial experience includes: Chief Informatics Officer and co-founder of Predictive Physiology and Medicine, Inc, Blooming, IN (2006-present); founder and interim CEO of MedeoLinx, LLC, Indianapolis, IN (2005-present); Head of Computational Proteomics and Principle Bioinformatics Scientist at Myriad Proteomics, Inc., Salt Lake City, UT (2002–2003); and Bioinformatics Computer Scientist at Affymetrix, Inc., Santa Clara, CA (1998–2002).

1 Biological data explosions and tools

Bioinformatics tools and systems perform a diverse range of functions on biological macromolecules including: data collection, data mining, data analysis, data management, and data integration. The earliest work in bioinformatics could date back to the first edition of the *Atlas of Protein Sequence and Structure*, compiled by Dayhoff et al. (1965). The *Atlas* later became the basis for the PIR protein sequence database (Wu et al., 2003). The term 'bioinformatics', and the practice of bioinformatics as a discipline did not emerge until the last 15 years. It arose from the recognition that efficient computational techniques were needed to study the huge amount of biological sequence information that was becoming available.

Since the first efforts of Maxam and Gilbert (1977) and Sanger et al. (1977), the DNA sequence databases have been doubling in size every 18 months or so. This trend continues unabated. Clearly, we have reached a point where correlated improvement in computer software and hardware are essential for the storage, retrieval, and analysis of biological sequence data. The sheer volume of data made it hard to find sequences of interest in each release of sequence databases, often represented and distributed as a collection of flat files in the early days.

The application of relational database management systems in biology partially addressed but did not solve the problem of managing biological sequence data. In 1998, a special issue of *Nucleic Acids Research* listed 64 different databanks covering diverse areas of biological research, and the nucleotide sequence data alone at over 1 billion bases. It became increasingly obvious that both the size and heterogeneity of biological data make the issues of information representation, storage, structure, retrieval and interpretation challenging. There has also been a change in user community. In the mid 1980s, fetching a biological entry on a mainframe computer was an adventurous step that only few dared. Now, at the end of the 1990s, thousands of researchers make use of biological databanks on a daily basis to answer queries, e.g., to find sequences

similar to a newly sequenced gene, or to retrieve bibliographic references, or to investigate fundamental problems of modern biology (Koonin and Galperin, 1997). New technologies, of which the World Wide Web (WWW) has been the fundamental driving force, have made it possible to create a numerous databanks and crosslinks between databanks.

2 Need for ontologies

Most biological public databases, until recently, still distribute their bulk contents as flat files. In some cases, indices were used for rapid data retrieval. In principle, all flat file formats are based on the organisational hierarchy of database, entry, and record. Entries are the fundamental entities of molecular databases, but in contrast to the situation in the living cell that they purport to describe, database entries store objects in the form of atomic, isolated, non-hierarchical structures. Different databases may describe different aspects of the same biological unit, e.g., the nucleic acid and amino acid sequences of a gene, and the relationship between them must be established by links that are not intrinsically part of the data archives themselves.

The development of individual databases has generated a large variety of formats in their implementations. There is consensus that a common language, or at least that mutual intelligibility, would be beneficial to end users although difficult to achieve. Attempts to unify data formats have included application of Backus–Naur based syntax (George et al., 1987), the development of an object-oriented database definition language (George et al., 1993) and the use of Abstract Syntax Notation 1 (Ohkawa et al., 1995; Ostell, 1990). None of these approaches has achieved the expected degree of acceptance. How to address the mechanisms of intercommunication between databases of different structure and format arrives the need for common semantic standards and controlled vocabulary in annotations (Pongor, 1998; Rawlings, 1998). This problem is especially prominent in comparative genomics. From a technological perspective, inter-genome comparisons rely on inter-database comparisons, which requires that the databases to be compared talk to each other in the same language: keywords, information fields, weight factors, object catalogues, etc.

The use of data standardisation could be addressed more effectively in the context of a more general logical structure – ontology. As noted by Hafner and Fridman (1996), general biological data resources are databases rather than knowledge bases: they describe miscellaneous objects according to the database schema, but no representation of general concepts and their relationships is given. Schulze-Kremer (1998) addressed this problem by developing ontologies for knowledge sharing in molecular biology. He proposed to create a repository of terms and concepts relevant to molecular biology, hierarchically organised by means of 'is a subset of ' and 'is member of' operators.

3 Biomedical ontologies

Existing traditional approaches do not address the complex issues of biological data discussed in earlier sections. However, recent work on ontologies intends to provide solutions to these issues. The term ontology is originally a philosophical term referred as '*the object of existence*'. Computer Science community borrowed the term

ontology to refer to a 'specification of conceptualisation' for knowledge sharing in artificial intelligence (Gruber, 1993). Ontologies provide a conceptual framework for a structured representation of the meaning, through a common vocabulary, on a given domain – in this case, biological or medical – that can be used by either humans or automated software agents on a the domain. This shared vocabulary usually includes concepts, relationships between concepts, definitions for these concepts and relationships and also the possibility of defining ontology rules and axioms; in order to define a mechanism to control the objects that can be introduced in the ontology and to apply logical inference. Ontologies in biomedicine have emerged because of the need for common language for effective communication across diverse sources of biological data and knowledge.

Several Biomedical Ontologies like UMLS (Baclawski et al., 2000) Gene Ontology (Ashburner et al., 2001), Protein Ontology (Sidhu et al., 2005), MGED Ontology (Whetzel et al., 2006), and TAMBIS Ontology (Baker et al., 1999) have developed, often reflecting mere relations of 'association' between what are called 'concepts', and serving primarily the purposes of information extraction from on-line biomedical literature and databases. In recent years, we have learned a great deal about the criteria, which must be satisfied if ontology is to allow true information integration and automatic reasoning across data and information derived from different sources.

4 Contributing papers

Ontologies for Biomedical Systems 2006 is the first special track dedicated to Biomedical Ontologies and Systems held at the *19th International Symposium on Computer-Based Medical Systems*. The goal of this track is to survey existing biomedical ontologies and reform them in such a way as to allow true information integration in biomedical domain. Authors are invited to submit original papers exploring the theories, techniques, and applications of biomedical ontologies. Only eight papers of the 17 submitted papers were accepted for oral presentation at the track. We selected high quality papers of the papers that were presented at CBMS 2006 special track and also published an open call for papers for the this special issue as well. We received 24 high quality submissions for this special issue. Each of the papers went through reviews by two experts in the field of biomedical ontologies, before we accepted eight papers for our special issue on Ontologies for Bioinformatics.

The selection of papers for this special issue discusses use of ontologies in various areas of bioinformatics. Now we briefly discuss contents of the contributing papers. In this special issue Digiampietri et al. (2007) propose a ontology based framework for bioinformatics workflows to support the specification and annotation of bioinformatics workflows, and to serve as the basis for tracking data provenance. Moreover, it uses techniques to support automatic or interactive composition of tasks. On the other hand another paper in this issue by Dhanapalan and Chen (2007) study various available semantic web technologies for integrating protein interaction data and describe an ontology-driven semantic data integration approach to address the weaknesses of the related approaches. Paper by Wolstencroft et al. (2007) in the issue explores issues in the development of the myGrid ontology, which is an OWL ontology designed to support service discovery through service annotation. Kupfer et al. (2007) describe database ontology for signal transduction pathways in their paper. Also a paper in this issue by

Dinakarpandian et al. (2007) discusses an approach for mapping Open Biomedical Ontologies by analysing aspects of overlapping relationships between them and provides an interoperability framework, called InterOBO.

In this issue Elmasri et al. (2007) describe an approach for modelling concepts and database implementation of complex biological data. In their paper Witte and Kappler (2007) investigate a novel approach for providing access to biological knowledge by employing Description Logics (DL)-based queries made to formal ontologies that have been created using the results of text mining full-text research papers. They demonstrate the feasibility of this approach with a system targeting the protein mutation literature. Finally, in their paper Yoo et al. (2007) investigate if biomedical ontology improves biomedical literature clustering performance in terms of the effectiveness and the scalability. For this investigation, they perform a comprehensive comparison study of various document clustering approaches.

5 Summary

In this issue we present a collection of high quality papers that discuss various aspects of ontologies in bioinformatics: Biological Data Modelling, Biomedical Data Integration using Ontologies, Biomedical Ontology Design, Mapping Biomedical Ontologies, Semantic Interoperability, Query Methodologies and Performance Analysis. We hope to cover all the major research areas in biomedical ontologies through this special issue. We are organising special track at CBMS 2007 for second year, and we hope to bring you more interesting research in biomedical ontologies in Special Issue on Ontologies for Bioinformatics II for *International Journal of Bioinformatics Research and Applications (IJBRA)* in 2008 as well.

Acknowledgements

We would like to thank all the authors for their valuable contribution to this special issue of IJBRA. We would also like to thank speakers, attendees, and conference general chairs for contributing to the success of the special track on Ontologies for Biomedical Systems at CBMS 2006. We would specially like to acknowledge the support of IJBRA Editor-in-Chief Professor Yi Pan for bringing this special issue together. We particularly would like to thank experts in our review board for help in reviewing the submissions.

References

Ashburner, M., Ball, C.A., Blake, J.A., Butler, H., Cherry, J.C., Corradi, J. and Dolinski, K. (2001) 'Creating the gene ontology resource: design and implementation', *Genome Research*, Vol. 11, pp.1425–1433.

Baclawski, K., Cigna, J., Kokar, M.M., Magner, P. and Indurkhya, B. (2000) 'Knowledge representation and indexing using the unified medical language system', Presented at *Pacific Symposium on Biocomputing*, Honolulu, Hawaii.

Baker, P.G., Goble, C.A., Bechhofer, S., Paton, N.W., Stevens, R. and Brass, A. (1999) 'An ontology for bioinformatics applications', *Bioinformatics*, Vol. 15, pp.510–520.

Dayhoff, M.O., Eck, R.V., Chang, M.A. and Sochard, M.R. (1965) *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, USA.

Dhanapalan, L. and Chen, J.Y. (2007) 'A case study of integrating protein interaction data using semantic web technology', *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, Vol. 3, No. 3, pp.286–302.

Digiampietri, L.A., Pérez-Alcázar, J.J. and Medeiros, C.B. (2007) 'An ontology-based framework for bioinformatics workflows', *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, Vol. 3, No. 3, pp.268–285.

Dinakarpandian, D., Tong, T. and Lee, Y. (2007) 'A pragmatic approach to mapping the open biomedical ontologies', *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, Vol. 3, No. 3, pp.341–365.

Elmasri, R., Ji, F., Fu, J., Zhang, Y. and Raja, Z. (2007) 'Modelling concepts and database implementation techniques for complex biological data', *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, Vol. 3, No. 3, pp.366–388.

George, D.G., Mewes, H-W. and Kihara, H. (1987) 'A standardized format for sequence data exchange', *Protein Seq. Data Anal.*, Vol. 1, pp.27–29.

George, D.G., Orcutt, B.C., Mewes, H-W. and Tsugita, A. (1993) 'An object-oriented sequence database definition language (sddl)', *Protein Seq. Data Anal.*, Vol. 5, pp.357–399.

Gruber, T.R. (1993) 'A translation approach to portable ontology specifications', *Knowledge Acquisition*, Vol. 5, pp.199–220.

Hafner, C.D. and Fridman, N. (1996) 'Ontological foundations for biology knowledge models', Presented at *4th International Conference on Intelligent Systems for Molecular Biology*, St. Louis.

Koonin, E.V. and Galperin, M.Y. (1997) 'Prokaryotic genomes: the emerging paradigm of genome-based microbiology', *Current Opinions in Genetic Development*, Vol. 7, pp.757–763.

Kupfer, A., Eckstein, S., Störmann, B. and Mathiak, B. (2007) 'A database ontology for signal transduction pathways', *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, Vol. 3, No. 3, pp.326–340.

Maxam, A.M. and Gilbert, W. (1977) 'A new method for sequencing DNA', *Proceedings of National Academic of Science*, Vol. 74, pp.560–564.

Ohkawa, H., Ostell, J. and Bryant, S. (1995) 'MMDB: an ASN.1 specification for macromolecular structure', Presented at *3rd International Conference on Intelligent Systems for Molecular Biology*, Cambridge, United Kingdom.

Ostell, J. (1990) *GenInfo ASN.1 Syntax: Sequences*, Technical Report 1, National Center for Biotechnology Information.

Pongor, S. (1998) 'Novel databases for molecular biology', *Nature*, Vol. 332, p.24.

Rawlings, C.J. (1998) 'Designing databases for molecular biology', *Nature*, Vol. 334, pp.447–447.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of National Academic of Science*, Vol. 74, pp.5463–5467.

Schulze-Kremer, S. (1998) 'Ontologies for molecular biology', Presented at *Pacific Symposium of Biocomputing*, Hawaii.

Sidhu, A.S., Dillon, T.S. and Chang, E. (2005) 'Ontological foundation for protein data models', Presented at *1st IFIP WG 2.12 & WG 12.4 International Workshop on Web Semantics (SWWS 2005), in Conjunction with on the Move Federated Conferences (OTM 2005)*, Agia Napa, Cyprus.

Whetzel, P.L., Parkinson, H., Causton, H.C., Fan, L., Fostel, J., Fragoso, G., Game, L., Heiskanen, M., Morrison, N., Rocca-Serra, P., Sansone, S., Taylor, C., White, J. and Stoeckert, C.J. (2006) 'The MGED ontology: a resource for semantics-based description of microarray experiments', *Bioinformatics*, Vol. 22, pp.866–873.

Witte, R., Kappler, T. and Baker, C.J.O. (2007) 'Enhanced semantic access to the protein engineering literature using ontologies populated by text mining', *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, Vol. 3, No. 3, pp.389–413.

Wolstencroft, K., Alper, P., Hull, D., Wroe, C., Lord, P.W., Stevens, R.D. and Goble, C.A. (2007) 'The myGrid ontology: bioinformatics service discovery', *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, Vol. 3, No. 3, pp.303–325.

Wu, C.H., Yeh, L.S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E., Vinayaka, C.R., Zhang, J. and Barker, W.C. (2003) 'The protein information resource', *Nucleic Acids Research*, Vol. 31, pp.345–347.

Yoo, I., Hu, X. and Song, I-Y. (2007) 'Biomedical ontology improves biomedical literature clustering performance: a comparison study', *Special Issue on Ontologies for Bioinformatics for International Journal of Bioinformatics Research and Applications (IJBRA)*, Vol. 3, No. 3, pp.414–428.