

Tutors: Vladimir Brusic and Judice Koh

Vladimir Brusic is a principal investigator at the Institute for Infocomm Research (I²R, former KRDL, LIT) Singapore, since 1998. He is the head of the Biodiscovery Group at I²R and an Adjunct Associate Professor of Bioinformatics and Immunology at the University of Canberra, Australia. Previously he had been a senior programmer at the Walter and Eliza Hall Institute, Melbourne, Australia. He received the PhD degree in Bioinformatics, and Master degrees in Information Technology, in Biomedical Engineering, and in Business Administration. His research interests include complex systems analysis, computer modelling of biological systems, and knowledge discovery from biological databases. He produced several specialist biological databases in the fields of molecular immunology and molecular venom science. He gave six invited tutorials including ISMB-98. He authored more than 60 scientific publications. Computational models of immune interactions that he developed have been successfully applied to the prediction and identification of vaccine targets in autoimmunity, cancer, and malaria research.

Judice Koh MTEch (Software Engineering) is a researcher at the Institute for Infocomm Research (I²R, former KRDL, LIT) Singapore and a PhD student at the National University of Singapore. Her research interests include biological databases, data warehousing, and data cleaning. She produced several specialist biological databases in the fields of molecular immunology and molecular venom science. She authored six scientific publications in the field of bioinformatics and biological databases. She designed, with V. Brusic, BioWare - a data warehousing system for rapid building of specialist biological databases.

References

11. Schönbach C., **Koh J.L.Y.**, Flower D.R., Wong L. and **Brusic V.** (2002). FIMM, a database of functional molecular immunology - update 2001. *Nucleic Acids Research*, 30, 226-229.
10. Srinivasan K.N., Gopalakrishnakone P., Tan P.T., Chew K.C., Cheng B., Kini R.M., **Koh J.L.Y.**, Seah S.H. and **Brusic V.** (2002). SCORPION, a molecular database of scorpion toxins. *Toxicon*, 40, 23-31.
9. Schönbach C., Kowalski-Saunders P. and **Brusic V.** (2000). Data warehousing in molecular biology. *Briefings in Bioinformatics* 1(2), 190-198.
8. **Brusic V.**, Zeleznikow J. and Petrovsky N. (2000). Molecular immunology databases and data repositories. *Journal of Immunological Methods* 238(1-2), 17-28.
7. Schönbach C., **Koh J.L.Y.**, Sheng X., Wong L. and **Brusic V.** (2000). FIMM, a database of functional molecular immunology. *Nucleic Acids Research* 28(1), 222-224.
6. **Brusic V.** and Zeleznikow J. (1999). Knowledge Discovery and Data Mining in

Biological Databases. *Knowledge Engineering Review* 14(3), 257-277.

5. **Brusic V.**, Zeleznikow J., Sturniolo T., Bono E. and Hammer J. (1999). Data cleansing for computer models: a case study from immunology. *Proceedings of ICONIP99 Sixth International Conference on Neural Information Processing*, IEEE, 603-609.
4. Hon L., Abernethy N.F., **Brusic V.**, Chai J. and Altman R. (1998). MHCWeb: Converting a WWW database into a knowledge-based collaborative environment. *Proceedings of AMIA Symposium*, 947-951.
3. **Brusic V.**, Wilkins J.S, Stanyon C.A. and Zeleznikow J. (1998). Data learning: understanding biological data. In Merrill G. and Pathak D.K. (eds.) *Knowledge Sharing Across Biological and Medical Knowledge Based Systems: Papers from the 1998 AAAI Workshop* pp. 12-19. *AAAI Technical Report WS-98-04*. AAAI Press.
2. **Brusic V.**, Rudy G. and Harrison L.C. (1998). MHCPEP - a database of MHC-binding peptides: update 1997. *Nucleic Acids Research*, 26, 368-371.
1. **Brusic V.**, Rudy G. and Harrison L.C. (1994). MHCPEP - a database of MHC binding peptides. *Nucleic Acids Research*, 22, 3663-3665.

Title and expected goals, objectives and motivation of the tutorial

Title: **Data Warehousing in Molecular Biology**

Intended audience:

Biologists and Computer Scientists interested in discovering new biological knowledge from biological databases through a three stage process: a) extraction of biological data from multiple sources (literature, public databases, and local experimental data, b) conversion and manipulation of that data, and c) knowledge discovery and data mining techniques. The audience is expected to have basic knowledge of biological databases (e.g. GenBank, SWISS-PROT) and database search techniques (BLAST). The tutorial will be structured to contain basic examples suitable for beginners followed by selected advanced database and algorithm concepts for more advanced bioinformatics users and developers. The tutorial will also contain a demonstration of software used for building specialist data warehouses, which will demonstrate the building of a small data warehouse from on-line sources and literature.

Detailed outline of the presentation.

Summary

A basic question for the biologist is “There are so many data out there. How do I access this data for analysis and extraction of relevant information?” Data warehousing takes the question one step further “How do I organize data about a particular subject so that I can efficiently perform complex analysis or mine these data?”

Data integration is a prerequisite for improving the efficacy of extraction and

analysis of biological information, particularly for knowledge discovery, and research planning. The data warehousing approach has been successfully applied for similar purposes in health care and chemi-informatics. Data warehousing has occasionally been used in molecular biology for creating subsets within existing databases. Common data warehousing goals are the formulation of querying, reporting, and complex analysis (multidimensional analysis and data mining). A data warehouse is structured to assist analytical tasks, rather than operational purposes (e.g. real-time data) and is therefore suitable for bioinformatics applications. The fuzziness and complexity of biological data represent major challenges in molecular biology data warehousing, and require high-level expert interpretations to suffice the requirements of the biological R&D arena. A molecular biology data warehouse as a subject-oriented, integrated, non-volatile, expert-interpreted collection of biological data in support of biological data analysis and knowledge discovery.

The main components of a data warehouse are target data, archive data, and metadata. The operations include transformation, extraction, monitoring, archiving, and purging. Access to diverse biological databases and efficient analyses of these data is an important factor for interpretation of experimental results, target identification, and research planning in molecular biology. However, the distributed nature of biological databases makes the access to and extraction of useful information complicated. Databases in molecular biology are characterized by various degrees of heterogeneity. These databases utilize different views of the domain, data formats, database management systems, and data manipulation languages. They also encode data at various levels of complexity. General-purpose sequence databases such as GenBank focus on the expansion and dissemination of information, and provide basic annotations. SWISS-PROT is a slow-growing, low-redundancy general database of protein sequences. It provides extensive annotations of functional aspects of proteins using a controlled terminology. These general sequence databases do not fit the definition of data warehouses, because they are primarily sequence repositories that are not subject-oriented. A specialized database has a narrower scope and therefore is more closely related to the data warehouse concept. We will discuss several data warehouse applications.

The critical issues in data warehouse design are data modeling, data transformation, data quality and annotation, automation of data warehousing process, integration of analysis tools, and implementation. We will demonstrate the BioWare software for building small specialise molecular data warehouses focusing on the discovery of protein function and identification of molecular targets of active proteins.

The tutorial will include specific examples. This is the overview of the data warehousing tutorial:

1. Introduction
 - Background
 - Definitions
 - Related work
 - Theoretical and practical problems
2. Biological Databases

- Primary and secondary molecular databases
 - Generalist vs. specialist databases
 - Data format and data exchange
3. Biological Data
 - complexity and hierarchical nature of processes that generate biological data,
 - fuzziness of biological data
 - biases and potential misconceptions in data
 - effects of noise and errors
 4. Data Integration
 - integrative systems
 - template-based systems
 5. Data cleaning
 - automated systems
 - manual systems
 6. Data mining techniques
 - knowledge discovery from databases
 - data mining function and algorithms
 - interpretation
 7. Conclusion
 - data warehousing for specialist systems
 - large-scale data warehouses
 8. Demonstration – BioWare
 - data acquisition
 - data preparation
 - warehouse implementation
 - data mining

References to be provided with the tutorials

- Koh J. and Brusica V. BioWare – A Framework for Data Warehousing in Molecular Biology. (In preparation).
- Schönbach C., Kowalski-Saunders P. and Brusica V. (2000). Data warehousing in molecular biology. *Briefings in Bioinformatics* 1(2), 190-198.
- Brusica V. and Zeleznikow J. (1999). Knowledge Discovery and Data Mining in Biological Databases. *Knowledge Engineering Review* 14(3), 257-277.
- Brusica V., Wilkins J.S, Stanyon C.A. and Zeleznikow J. (1998). Data learning: understanding biological data. In Merrill G. and Pathak D.K. (eds.) Knowledge Sharing Across Biological and Medical Knowledge Based Systems: Papers from the 1998 AAAI Workshop pp. 12-19. *AAAI Technical Report WS-98-04*. AAAI Press.