SLAD – An Integrated Swine MHC Database: A Case Study of Data Warehousing in Molecular Immunology

Christian Schönbach schoen@krdl.org.sg
Vladimir Brusic vladimir@krdl.org.sg
Judice L.Y. Koh judice@krdl.org.sg
Computational Immunology Group, Bioinformatics Center Research Unit
Kent Ridge Digital Labs, Singapore 119613, Singapore

1 Introduction

"Outcomes Research" in molecular immunology is driven by faster, cheaper and increasingly sophisticated methods such as miniaturisation, automation, and data integration. The latter is a prerequisite for efficient information analysis, knowledge discovery, and eventually research planning. The data warehousing approach has been successfully applied for managing clinical data [1], but rarely in exploratory biological research. In order to clarify aspects of data warehousing in molecular immunology we constructed a small model database of swine major histocompatibility antigens (swine MHC or SLA) using warehousing principles.

2 Method

2.1 Dimensional modeling

Dimensional modeling [2] is helping to conceptualise and visualise data models. During the modeling process of SLAD four dimensions were defined: sequence analysis, sequence information, literature, and physical maps.

2.2 Data transformation

Data were acquired from Entrez, Genbank [3], and printed articles. Data transformation included several steps: a) raw data downloading, b) data standardisation, c) data cleaning, d) data annotation, and e) consistency check. Performing these steps required a high level of domain expertise to ensure acceptable accuracy and quality of resulting target data.

3 Results and Discussion

SLAD cannot be considered as a data warehouse because of limited integration capability and small size. However our experience showed that dimensional modeling is most appropriate in an environment where computer scientists define requirements for database building and biologists decide how much and which proportion of data is necessary and sufficient for performing biological analysis. Great effort had to be spent on accessing and transforming data sources. A surprisingly large number of errors were detected in the raw data. In 163 published data sources 36 errors were found for example, conflicting tissue types, gene and locus names, erroneous translation and non-functional Entrez links. Due to the complexity of errors, elimination of erroneous data requires extensive knowledge of the domain and renders automation a difficult task.

4 Accessibility

The SLA Database (SLAD) is accessible via the URL http://charon.ima.org.sg/slad/

References

[1] Geisler, M.A. and Will, D., Implementing enterprisewide databases: a challenge that can be overcome, *Top. Health Inf. Manage.*, 19:11–18, 1998.

[2] Ballard, C., Herreman, D., Schau, D., Bell, R., Kim, E., and Valencic, A., Data modeling techniques for data warehousing, *IBM International Technical Support Organization*, 42–47, 1998.

[3] Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A., and Wheeler, D.L., GenBank, *Nucleic Acids Res.*, 27:12–17, 1999.