OODB4Genomics: An object-oriented database approach for biomedical data in clinical bioinformatics

Rainer von den Berken[1], Markus Gumbel[1]

[1] Mannheim University of Applied Sciences, Institute for Medical Informatics

Contact m.gumbel@hs-mannheim.de

Introduction: Bioinformatics methods are more and more used in clinical medicine. Examples are the support for diagnoses or pharmacogenomics including personalized medicine (overview in [1]). There are plenty of databases with molecular-biological content available. However, databases or information systems with a focus on molecular medicine that also consider individual patients are rare. Thus, a universal data (or object) model, which is a must for any medical information system, is not available. Instead, molecular entities are typically modeled with ontologies like GO [2], markup languages like SMBL [3] or GSVML [4], terminologies like UMLS [5], or with HL7s “Clinical Genomics” domain model [6]. Software like BioMart [7] enables the integration and querying of multiple data sources. However, all these diverse and independent approaches still have interoperability issues [8]. What if we circumvent these downsides by creating a uniform object-oriented (OO) model for molecular medicine before we try to integrate? We have analyzed the pros and cons of an object-oriented domain model and applied it to a prototype database.

Material and Methods: A review of the above mentioned methods and techniques for modeling molecular-biology entities was conducted. The outcome was used as input for the domain model. All classes are modeled in the Unified Modeling Language (UML) [9]. Scala [10] is used together with the OO database db4o [11] for the prototype. Publicly available data of the pharmacogenomics knowledge base (PharmGKB) [12] was used to set up a database with real-world data as a proof of concept.

Results: We have successfully developed a domain model and a database which is based on this model. Our approach uses several design patterns to separate generic classes from data-source-dependent classes (e.g. PharmGKB). The model consists of 21 major generic classes like Gene or Drug and 8 major PharmGKB-dependent classes like PKGBGene or PGKBDrug. The database contains 26,216 genes, 3,196 diseases and 2,952 drugs and supports patient-related polymorphism (SNPs). Db4o’s powerful intrinsic query methods enable a wide variety of queries and the DB could easily be extended to a medical information system. Test queries indicate no significant performance problems as expected [13].

Discussion: Little schemas or class models of existing databases are publicly available. Our perception is that each DB scheme is used only internally and serves its primary purpose. The feasibility of a uniform class model for molecular biology is still under discussion [14]. However, we found that the molecular biology/molecular medicine domain can naturally be described as objects. We believe that the OO approach fits nicely for molecular networks where a relational DB approach would require many joins (see also [15]). Future work will incorporate support for copy number variations and will analyze how queries can be extended with lazy loading mechanism (also over a web service).

References

[1] Coleman, W. B. and Tsongalis, G. J. (eds.): Molecular Diagnostics. Humana Press 2010

[2] Ashburner et al.: Creating the Gene Ontology Resource: Design and Implementation. Genome Research, 2001, Vol. 11, pp. 1425-1433

[3] Hucka, M. et al.: The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core. Nature Precedings, 2010

[4] Nakaya, J. et al.: Genomic Sequence Variation Markup Language (GSVML). Int J Med Inform, 2009, 79, pp. 130-142

[5] Bodenreider, O. et al: Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp. 1998, pp. 815-9.

[6] HL7v3 Clinical Genomics. http://www.hl7.org/v3ballot/html/domains/uvcg/uvcg.htm (last access 22.03.2011)

[7] Smedley, D. et al.: BioMart - biological queries made easy. BMC Genomics. 2009, pp. 10-22

[8] Katayama, T. et al: The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. J. Biomedical Semantics, 2010, Vol. 1

[9] UML specification of the Object Management Group: http://www.omg.org/spec/UML/2.3/ (last access 13.07.2011)

[10] Odersky, M. et al.: Programming in Scala. Artima Press 2008

[11] Paterson, J., Edlich, S.: The Definitive Guide to db4o. Apress 2006.

[12] Thorn, C.F. et al: Pharmacogenomics and bioinformatics: PharmGKB. Pharmacogenomics, 2010, Vol. 11(4), pp. 501-505

[13] van Zyl, P. et al.: Comparing the Performance of Object Databases and ORM Tools. Proceedings of SAICSIT, 2006, pp. 1-11

[14] Birney, E. and Clamp, M.: Biological database design and implementation. Briefings in Bioinformatics. 2004. Vol. 5. pp. 31-38

[15] Ireland, C. et al.: A Classification of Object-Relational Impedance Mismatch. IEEE Computer Society: Advances in Databases, First International Conference, 2009, pp. 36-43