COMPARISON BETWEEN RELATIONAL DATABASE AND XML IN QUERYING MOTIF SEQUENCE

MOHD TAUFIK BIN MISHAN

UNIVERSITI TEKNOLOGI MALAYSIA

COMPARISON BETWEEN RELATIONAL DATABASE AND XML IN QUERYING MOTIF SEQUENCE

MOHD TAUFIK BIN MISHAN

A project report submitted in partial fulfillment of

the requirement for the award of the degree of

Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

OCTOBER 2009

ABSTRACT

An enterprise information approach is the process of building models which include with process models, data models and resource models. An enterprise in general is a unit of economic organization or activity which these activities are required to develop and deliver products or services to a customer. An enterprise also includes a number of functions and operations. The aim of this research is to study an enterprise approach and their appropriate model to integrate multiple biological data use by the scientist in protein secondary structure prediction process. This project needs to investigates the data format and database will be use by the scientist at the Bioinformatics Research Lab, FBB, UTM in order to predict protein secondary structure process. The size, complexity and number of database used for predict protein secondary structure process and integrating all the data from the different database into one database is a challenging problem. This project approach have integrated different database such as Prosite, Blast, Prints and PDB and transformed these databases in flat file format and other format into relational form using XML and asp.net. As a result, this project showed some tool can search different data and different sizes of protein secondary structure data stored in the relational database and the result can be retrieved faster and reliable compared to XQuery direct from the XML file. A prototype web based user interface is provided to allow user access and search for protein secondary structure prediction in repository and local relational database.

ABSTRAK

Enterprise information adalah satu pendekatan dalam membina model yang merangkumi pemprosesan model, model data dan sumber model.Pada amnya, enterprise adalah terdiri daripada satu unit organisasi ekonomi ataupun aktiviti yang diperlukan untuk membangunkan dan menyampaikan produk, perkhidmatan dan juga operasi. Secara keseluruhannya, kajian ini adalah untuk mengkaji kaedah enterprise dan model yang bersesuaian bagi menyambung dan menggabungkan kepelbagaian data biologi dimana ianya digunakan oleh para penyelidik dalam proses peramalan "*protein secondary structure"*. Kajian ini memerlukan penyiasatan terhadap format data dan, pangkalan data yang akan digunakan oleh penyelidik di Bioinformatics Research Lab, FBB, UTM bagi membolehkan usaha proses peramalan "*protein secondary structure*" ini dilakukan. Saiz, kerumitan dan bilangan pangkalan data yang digunakan dalam proses ini merupakan satu masalah yang mencabar. Pendekatan yang digunakan dalam projek ini adalah dengan mengabungkan pelbagai data daripada pangkalan data yang berbeza seperti pangkalan data daripada Prosite, Blast, Prints dan PDB yang mana data-data berformat "flat file" daripada pangkalan data yang berbeza tersebut diubah kepada format yang berkaitan dengan menggunakan XML dan asp.net. Sebagai keputusannye, projek ini menunjukkan satu peralatan yang boleh digunakan untuk mencari data "*protein secondary structure*" yang berbeza dan saiz data "*protein secondary structure*" yang mana ianya telah disimpan di dalam satu pengkalan data. Keputusan boleh dicapai dan lebih boleh dipercayai di bandingkan dengan mencari data dengan menggunakan Xquery terus daripada fail XML. Satu prototaip antaramuka pengguna dihasilkan bagi membolehkan pengguna mencapai dan mencari data "*protein secondary structure*" di dalam gudang (repository) dan pangkalan data relational.

TABLE OF CONTENT

CHAPTER 1

INTRODUCTION

1.1 Introduction

The Protein structure prediction is a relatively young area of bioinformatics research, but a fast growing one. Despite this, modeling and prediction tools have already been built specifically for novice users to produce approximate models of the three-dimensional structures of protein gene products, solely from their newly acquired nucleotide sequence. The result of protein structure prediction will be use to understand the relationship between motif sequence, structure and function of protein. Before that scientist need to search for motif sequence from the bioinformatics database, after get the result of motif sequence, scientist need to search for assigned secondary structure to get result of motif sequences with secondary structure assignment from the different bioinformatics database or another bioinformatics data sources. This task show that the scientist often to retrieve data from multiple biological data source to solve their research problem and understanding the relationship between motif sequence, structure and function. They vary type of stored data, data format, and access methods. In addition, there is a terminology discrepancy at the data level and at the schema level, which even more

complicates the data retrieval process. Scientist needs to decide which data source to access and in which order, how to retrieve the data and how to combine the results. The task of retrieving the data requires a great deal of effort and expertise on the part of the scientist. The scientists also have to take into account at bioinformatics where data source schemas change and new data source are developed. Besides that, scientists also waste their time to search the data from the different data sources.

1.2 Problem background

Protein structure prediction is the process which requires scientists to perform a variety of searching motif sequence and secondary structure procedures in order to answers their analytical result. To perform these procedures, scientists need to use variety types of sequence motif data from databases which are available from bioinformatics web sites. Scientists need to query motif sequences with secondary structure assignments simultaneously. Scientist needs to search for the motif sequence first to get the result of motif sequence. To perform this searching, scientist need to use two different website. The example of website being used by scientist are MOTIF (http://motif.genoma.jp/MOTIF.html), PROSITE (http://www.expasy.ch/prosite/) is a database of sequences characteristic of protein motifs (fragments of protein sequence known to be associated with a particular structure or function). After get the result of motif sequence scientist need to used BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) searches for homology between DNA or protein sequences but in this case scientist used this website to search for assigned secondary structure and get the result of motif sequence with secondary structure assignments. Scientists can visualization of protein structure by view alignment and view structure. Scientist used Motif 3d viewer application example like jmol, Web Mol applets, KiNG and webMol which requires java-enables browsers to view the protein structure. Next step scientist need to locating the position of motif within the global structure. To perform this task scientist need to

use a PDB website (http://www.rcsb.org). Scientist need to search motif in PDB and the scientist will be get the result and understand the relationship between motif sequence, structure and function. Based on the overall process it show that scientist need to searching which requires moving their query from one bioinformatics database to another. Result of one process/ from one website become input to another process or website it make in accuracy of experiment result. Scientist also needs to query motif sequence with secondary structure assignment simultaneously but currently there is no bioinformatics application to do that task.

1.3 Problem statement

How the enterprise approach can integrate and model multiple biological data for query searching motif sequence in protein secondary structure prediction process?

i) What is enterprise information approach?
ii) What is protein secondary structure?
iii) What are the problem integrating multiple biological data?
iv) What is microarray analysis process?
v) What are the data uses for this project?
vi) How they share data?
vii) What is model use in enterprise approach?
viii) What is XML?

1.4 Project aim

This project aim to study the appropriate method to integrate multiple biological data use by the scientist in protein secondary structure prediction process. This project need to investigate s the data format will be use by the scientist at the Bioinformatics Research Lab, FBB, UTM in order to predict protein secondary structure process. Besides that, this study also investigates the data model which can be use to predict protein secondary structure and compare among the data model which data model can be use effectively to predict the protein function process.

1.5 Objective

In order to accomplish the aim of the study, few objectives have been identified as stated below.

1. Identify the database and the data which currently use in protein secondary structure prediction process.
2. Investigating the enterprise based approach for data integration, especially for relational database in data integration.
3. Implement the relational based data model for interfacing multiple biological data.

1.6 Project scope

The main focus of this study is to minimize the problem of searching the motif sequence and protein secondary structure at the same time. The scopes of this project are as follows:

1. This study focuses on the work process of protein secondary structure prediction only.
2. The focus the work process is in dry experiment (bioinformatics experiment).

1.7 Significant of the Study

This study evaluates the performance of relational metadata approach that will be used to provide with model to integrate the multiple biological data. Compare the approach with the other approach common be used for integrate the multiple biological data resources in order to solve the problem of this project. The result of the study is contributed to the identification of new learning method. This new approach could be used to the development of a methodology that will be of value in future studies of bioinformatics improvement.

1.8 Organization of the report

This report consists of five chapters. The first chapter presents introduction to the project and the background of problem on why is the study is being conducted. It also gives the objectives and scope of the study as well as the significance of the project. Chapter 2 reviews on introduction of the protein structure, database of protein structure, The method that can be use to predict the protein structure error function. Profile analysis of the protein structure also discusses in this chapter. Chapter 3 discusses on the project methodology used in the study. It explains details the method that can be use for this project. Chapter 4 is the experimental result and discussion. Chapter 5 is the conclusion and suggestions for future work.

REFERENCES

Aranow E.,(1991). "Modeling Exercises Shape Up Enterprises". In: Software Magazine Vol.11 , p. 36-43

Barrett Tanya, Dennis B. Troup,Stephen E. Wilhite, Pierre Ledoux,Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva,Maxim Tomashevsky and Ron Edgar.,(2005) *NCBI GEO: mining millions of expression profiles--database and tools.*Nucleic Acids Res. **33**(Database issue): p. D562-6.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. Scientific American, 284(5), 34-43.

Booch and Rumbaugh (1995), "Unified Method for Object-Oriented Development". Grady Booch y James Rumbaugh. Documentation Set V0.8. Rational Software Corporation.

Bray T, Paoli J, Sperberg-McQueen CM, Maler E, Yergeau F (2004): Extensible Markup Language (XML) 1.0 (Third Edition) W3C Recommendation 04 February 2004. *W3C Recommendation* [http:// www.w3.org/TR/2004/REC-xml-20040204].

Brazma Alvis. Pascal hingamp, John Quackenbush Gavin Sherlock, Paul Spellman, chris stoeckert, John Aach, Wilhelm Ansorge, Catherine A Ball, Hellen C. Caustan ., (2001)*Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nat Genet, 2001. **29**(4): p. 365-71.

Breitkreutz, B.J, Paul Jorgensen , Ashton Breitkreutz ' and Mike Tyers.,(2001) *AFM 4.0: a toolbox for DNA microarray analysis.* Genome Biol, **2**(8): p. SOFTWARE0001.

Casey. R, (2006). How Federated Databases Benefit Bioinformatics Research. Beyenetwork . US edition.

Codd, E. F (1970) : *A relational model for large shared databanks, in*: Comm. ACM, Bd. 13, Nr. 6, S. 377-387.

Connolly T.M and Begg C.E., (2005). *Database systems: A practical approach to design, implementation and management.* Fourth edition. England: Pearson Education Limited.

Darling, C. (1996*). How to integrate your data warehouse*. *Datamation* (May 15) 40–52.

David G Nohle and Leona W Ayers., (2005) The tissue microarray data exchange specification: A document type definition to validate and enhance XML data *BMC Medical Informatics and Decision Making* 2005, **5**:12

Demeter1 J, Beauheim2 C, Gollub3 J, Tina Hernandez-Boussard2 T.N, Jin1 H, Maier1 D, John Matese4 J.C, Nitzberg1 M, Wymore1 F, Zachariah1 Z.K, Patrick O. Brown1,5,Sherlock2 G and Catherine A. Ball1.,(2006) *The Stanford Microarray Database: implementation of new analysis tools and open source release of software*, Nucleic Acids Research, Vol. 00, Database issue D1–D5.

Dittrich K.R (1986) ``Object-Oriented Database System : The Notions and the issues", in : *Dittrich, K.R. and Dayal, U. (eds): Proceedings of the 1986 International Workshop on Object-Oriented Database Systems*, IEEE Computer Science Press

Fellenberg, K. ,Christian H Busold, Olaf Witt**,** Andrea Bauer**,** Boris Beckmann**,** Nicole C Hauser**,** Marcus Frohme**,** Stefan Winter **,** Jurgen Dippon and Jorg D Hoheise, ., *Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis.* Bioinformatics, 2002. **18**(3): p. 423-33.

Ghosh, D., *Object-oriented transcription factors database (ooTFD).* Nucleic Acids Res,2000. **28**(1): p. 308-10.

Golub, T.R., D. K. Slonim, P. Tamayo,1 C. Huard,1M. Gaasenbeek,1 J. P. Mesirov,1 H. Coller,1 M. L. Loh,2 J. R. Downing,3 M. A. Caligiuri,4 C. D. BloomÞeld,4 E. S. Lander1,5*.,(1999) *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 286(5439): p. 531-7.

Gomez-Perez A., Fernandez-Lopez , M.,Corcho O. (2003) *Ontological Engineering*. Springer.

Guarino N. (1998) Formal Ontology in Information Systems. First international conference on formal ontology in information systems, Italy, Ed. Guarino, 3-15.

Guenther Rand J.Radebaugh., (2001). *Understanding metadata: a guide for libraries*. NISO press. National Information Standards Organization.

Gruber .T. (1993). Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, special issue on Formal Ontology in Conceptual Analysis and Knowledge Representation. Eds, Guarino, N. & Poli , R.

Harold E. R, (2002)Means WS: *XML In A Nutshell* 2nd edition. O'Reilly & Associates,Inc.

Heflin J. (2004). *OWL Web Ontology Language Use Cases and Requirements*. W3C Recommendation.WWW.W3.org.

Hegde P, Rong Qi, Kristie Abernathy, Cheryl Gay, Sonia Dharap, Renee Gaspard, Julie Earle-Hughes, Erik Snesrud, Norman Lee, and John Quackenbush.., (2000) *A Concise Guide to cDNA Microarray Analysis – II*., Based upon Biotechniques, 29(3) The Institute for Genomic Research, Rockville.

HREF1

(http://www.freebmd.org.uk/Flat.html, 2009)

HREF 2

(http://en.wikipedia.org/wiki/Spreadsheet, 2009)

HREF 3

(http://www.stg.brown.edu/service/xmlvalid/., 2009)

Inmon B., (1993). *The Operational Data Store*. InfoDB,

Jason T. L. W, Mohammed J. Z, Hannu T. T. Toivonen, Dennis S., (2005) Data mining in bioinformatics, edition , : Springer

Jornsten1 R, Ouyang2 M and Wang3 H.Y Hegde P, Rong Qi, Abernathy K, Gay C, Dharap S, Gaspard R, Earle J., (2007)*A meta-data based method for DNA microarray imputation*, biomed Central.

Juanle .W and Y. songcai Y., (2004). *Study on Web-Oriented Geo-Data Sharing Infrastructure and Key Techniques Based on Metadata*. Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Science Beijing, China

Kim. (1999*). an object-oriented data model***.** Volume 17 , Issue 3.

Knudsen, T.B. and G.P. Daston, *MIAME guidelines.* Reprod Toxicol, 2005. **19**(3): p. 263.

Lassila, O. and McGuiness, D. (2001). The role of Frame-Based Representation on the Semantic Web. Technical Report KSL-01-02, Stanford, California.

Maurer M, Molidor R, Sturn A, Hartler J, Hackl H, Stocker G, Prokesch A, Scheideler M and Trajanoski Z., (2005) MARS: Microarray analysis, retrieval, and storage system. Biomed central.

Maziarz, M.**,** Clement Chung, Daniel J. Drucker, and Andrew Emili.,(2005) *Integrating global proteomic and genomic expression profiles generated from islet alpha cells: opportunities and challenges to deriving reliable biological inferences.* Mol Cell Proteomics, 2. **4**(4): p. 458-74.

Mustafa Atay,Artem Chebotko,Dapeng Liu,et al.(2006) Efficient schema-based XML-to-Relational data mapping[J].,Information Systems.

Ostie J.K., (1996). "An Introduction to Enterprise Modeling and Simulation"

Petrie Jr. C.J., (1992). *"Introduction", In: Enterprise Integration Modeling* - Proceedings of the First International Conference MIT Press, p. 563.

Rayner T.F, Rocca-Serra1 P, Spellman2 P.T, Helen C, Causton3, Farne1 A, Holloway1 E, Irizarry4 R.A, Liu5 J, Maier6 D.S, Miller7 M, Petersen8 K, Quackenbush9 J, Sherlock10 G, Stoeckert Jr5 C.J, White9 J, Patricia L Whetzel5, Wymore6 F, Parkinson1 H, Sarkans1 U, Ball6 C.A and Brazma*1 A., (2006) *A simple spreadsheet-based, MIAME-supportive format for data: MAGE-TAB*, Biomed central, BMC Bioinformatics.

Rui, H. and M.J. Lebaron, *Creating tissue microarrays by cutting-edge matrix assembly.*Expert Rev Med Devices, 2005. **2**(6): p. 673-80.

Schageman, J.J., M. Basit, T.D. Gallardo, H.R. Garner, and R.V. Shohet., (2002) *MarC-V: a spreadsheet-based tool for analysis, normalization, and visualization of single cDNA microarray experiments.* Biotechniques, **32**(2): p. 338-40, 342, 344.

Shaya E., Thomas B.,(2001) Specifics on a XML Data Format for Scientific Data, ASP Conference Series, Vol. 238,

Striebel, H.M., Birch-Hirschfeld E, Egerer R, Földes-Papp Z.,(2003) *Virus diagnostics on microarrays.* Curr Pharm Biotechnol, 2003. **4**(6): p. 401-15.

Tomlinson R.F., (2007). Thinking about GIS: geographic information system planning for manager. Third edition. ESRI Inc.,

Vaduva A and Dittrich K.R., (2001). Metadata Management for Data Warehousing: Between Vision and Reality. Database Engineering & Applications. IEEE explore.

Vernadat F.B., (1997). Enterprise Modelling Languages ICEIMT'97 Enterprise Integration - International Consensus. EI-IC ESPRIT Project 21.859.

William M. Shui,Franky Lam,Damien K. Fisher, (2005)Querying and Maintaining Ordered XML Data using Relational Databases[J].Conferences in Research and Practice in Information Technology, Vol.39.

Willy A. Valdivia G, Christopher D2., (2006) *MICROARRAY DATA MANAGEMENT An Enterprise Information Approach: Implementations and Challenges.* Orion Integrated Biosciences Inc. New York, USA.

Wu, C.H.**,** Hongzhan Huang, Leslie Arminski, Jorge Castro-Alvear, Yongxing Chen, Zhang-Zhi Hu, Robert S. Ledley, Kali C. Lewis, Hans-Werner Mewes[1], Bruce C. Orcutt, Baris E. Suzek, Akira Tsugita[2], C. R. Vinayaka, Lai-Su L. Yeh, Jian Zhang and Winona C. Barker .,(2002) *The Protein Information Resource: an integrated public resource of functional annotation of proteins.* Nucleic Acids Res,. 30(1): p. 35-7.

Xie Yi-wu, wang Chen-yang, Cao Zhi-ying ， Chen Yan et al., (2007) Research on Store XML Data in Relational Database Based on XML Schema IFIP International Conference on Network and Parallel Computing - Workshops

Yang Y.H and Speed. T., (2002). Dudoit, S. and T.P. Speed, *A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs.* Biostatistics, 2000. **1**(1): p. 1-26

Zeeberg, B.R., Weimin Feng[2]**,** Geoffrey Wang[3]**,** May D Wang[2]**,** Anthony T Fojo[1]**,** Margot Sunshine[4]**,** Sudarshan Narasimhan[4]**,** David W Kane[4]**,** William C Reinhold[1]**,** Samir Lababidi[1]**,** Kimberly J Bussey[1]**,** Joseph Riss[5]**,** J Carl Barrett[5] and John N Weinstein[1],(2003) *GoMiner: a resource for biological interpretation of genomic and proteomic data.* Genome Biol, . 4(4): p. R28.