31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A Method of Biological Pathway Similarity Search Using High Performance Computing

Keyuan Jiang, Yingmeng Huang, and Joseph Robertson

Abstract—**Comparative study of biological pathway structures and composition can aid us in elucidating the functions of newly discovered pathways, understanding evolutionary traits, and determining missing pathway elements. A method has been developed to perform pair-wise comparison and similarity search of biological pathways. The comparison determines the differences of each pair of pathways represented in the XML format. The similarity search uses a scoring mechanism to rank the similarities of the pathway in question against those in the pathway repository. To achieve a reasonably good performance, the method is being implemented using the Condor high performance computing environment.**

I. Introduction

Biological pathways and networks are an essential model representing our current understanding of living systems at the cellular and molecular levels, and play an important role in determining biological functions and helping us understand diseases and discover new drug targets [1]. Advances in laboratory technologies have resulted in an accumulation of a significant amount of pathway data, with more than 200 publicly accessible pathway databases available on the Web [2, 3], and more pathway data will become available in the years to come.

Biomedical discoveries have long been aided by comparing the biological molecules under study against those that are known to discern the potential functions of the molecules in question. Especially, the completion of the Human Genome Project and the availability of similarity search tools such as BLAST [4] have accelerated biomedical discoveries by studying biological molecules and processes at genomic scale.

Similarly, comparative study of biological pathways can also promote biomedical discoveries [5] and complement other biological research techniques. While the field of comparative pathway study is still at its early stage, several efforts have been made in biological discoveries, for example, to uncover conserved pathways within bacteria and yeasts [6, 7], to identify functionally related proteins [8], and to infer phylogenetic properties [9, 10]. Other potential applications of comparative pathway analysis include studying human pathway functions from known model organisms and discovering missing molecules in a particular pathway [5].

Despite these pioneering efforts, the existing tools for the comparative pathway study are limited in their capacities of comparison and database querying. Protein-protein interactions are the simplest form of biological pathway, and several tools were developed for comparing interactions in either the linear fashion [11] or the graph format [12-14], focusing primarily on proteins or enzymes. Although proteins play an important role in pathways such as signal transduction pathways, metabolic pathways also involve many non-protein metabolites. Tools primarily for comparing metabolic pathways were also developed [7, 15]. Forst et al. [9] argued that pathway comparison should include more than metabolite contents, suggesting that the structure of pathways should also be compared because the pathway structure represents the process/functional steps involved in the pathway. Due to the variations of biological pathways (such as metabolic pathways, signal transduction, protein-protein interaction network and regulatory pathways), there have been a number of data models for representing different pathways [16]. The existing tools are only capable of processing a single particular type of pathways. There is a lack of powerful tools capable of comparing sophisticated pathway compositions and topologies.

Although there exist more than 200 publicly accessible pathway databases, querying each one of them requires a different access method, and each database has its own data model, making it extremely difficult, if not impossible, to perform efficient queries against those pathway databases using a consistent method. Contrary to biological sequence repositories such as GenBank hosted at the National Center for Biotechnology Information of the National Library of Medicine) at NIH, EMBL (European Molecular Biology Laboratory), and DDBJ (the DNA Databank of Japan), there has been a lack of the similar effort to maintain and support a centralized pathway repository based upon a community-developed standard format. Many pathway comparison tools only have access to a limited set of pathways, potentially overlooking some important pathways relevant to the one being investigated.

Inspired by sequence comparison methods such as BLAST for sequence similarity searches, we are developing a similarity search system for biological pathways that is made up of 1) an XML-based pathway comparison method capable of processing various types of biological pathways, 2) a similarity scoring mechanism for ranking the pathway

This work was supported in part by a Summer Research grant from the U.S. Department of Energy/Northwest Indiana Computational Grid (DOE/NWICG).

K. Jiang, Y. Huang and J. Robertson are with the Department of Computer Information Technology and Graphics, Purdue University Calumet, Hammond, IN 46323 (tel: 219-989-2035; fax: 219-989-3187; e-mail: {jiang, huangy, robertjo}@calumet.purdue.edu).

similarities or differences, and 3) a high performance implementation of the pathway similarity search using the Condor high performance computing environment [41].

II. METHODS

A. The XML-based Pathway Comparison

To perform queries against pathway datasets, data have to be stored in a centralized repository and in a format consistent across various datasets and various pathway types including pathways, networks, and interactions. Unlike biological sequences which are linear in their primary structure, pathways are commonly represented in more sophisticated data structures such as graphs [16, 17]. Such a representation poses challenges in performing pathway comparisons at genomic scale.

First, until recently, there has been no agreed-upon standard data format among data providers and the research community, resulting in multiplicity and duplication of pathway datasets, each with its own data model and access methods. Second, due to the heterogeneity of the pathway data models, it is nearly impossible to develop a pathway comparison method that will work with various pathway data structures and formats. Each comparison algorithm is associated with a specific data structure. Having a consistent data model is a prerequisite for efficient pathway comparisons.

Several efforts have been made in order to unify the data format of pathways, notably BioPAX [18], PSI-MI (Proteomics Standard Initiative – Molecular Interactions) [29], SBML (Systems Biology Markup Language) [30], and CellML (Cell Markup Language) [31]. Among these formats, the BioPAX standard is considered to be the most expressive, capable of representing various types of interactions and pathways based upon different levels [32]. The BioPAX standard, an object-oriented representation, defines components of a pathway as classes (entities), and each class is made up of a collection of attributes. A class can represent a physical object such as an RNA molecule or a process in a pathway such as a biochemical reaction. Based upon its ontology, the BioPAX standard is a special XML format for representing biological pathways and their components, allowing each instance (or object) of a BioPAX class to be represented as an XML document. Currently, several major pathways hosts including BioCyc [19, 20], KEGG [21, 22], and Reactome [23-25], provide their datasets in the BioPAX format.

With the unified data format and the availability of datasets, we have developed an object model based repository to store pathway datasets in the native BioPAX format to facilitate the retrieval of pathways and/or their components from the datasets provided by different data hosts [26, 27]. In our approach, objects of each individual pathway are stored in the "native" XML datatype column to leverage the power of XQuery (XML Query) [28]. This pathway repository provides a foundation for developing a pathway similarity search tool which allows querying a large collection of pathways stored in an efficient object-model based data store. Currently, the pathway repository contains datasets of three organisms (Homo sapiens, E. coli K-12, E. coli O157:H7) from BioCyc and KEGG with a total of 594 pathways, 2,721 biochemical reactions and 4,223 catalyses. More datasets are being imported into the repository.

Our approach to pathway comparisons is to determine the structural and compositional differences and/or similarities of two given pathways. The XML representation of biological pathway datasets reduces the determination of structural and compositional differences of a pair of given pathways to the determination of the topological and compositional differences of two XML documents. XML differencing is the primary technique used to determine such differences. Several algorithms and implementations of XML differencing have been developed by others. They include x-diff [33], diffX [34], DiffXML [35], LogiLab's xmldiff [36], DecisionSoft's xmldiff tool [37], IBM's XML Diff and Merge Tool [38], and Microsoft's XML Diff and Patch Utility [39].

The Microsoft's XML Diff and Patch Utility was chosen in our implementation for the following reasons: (1) the availability of its source code which allows us to make

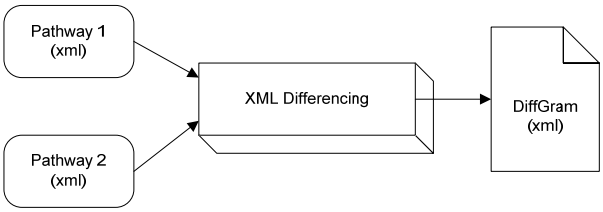

Fig. 1. Using the XML representation of biological pathways, the comparison of two pathways can be reduced to comparison of two XML documents.

modifications to suite our particular requirements, (2) the flexibility provided that allows to use different criteria for differencing XML documents, (3) the XML format of its output – the DiffGram format, and (4) easy implementation in our Condor high performance computing environment discussed later.

Before any pair of the native BioPAX compliant XML documents can be compared, they are preprocessed in order to (1) filter out insignificant pathway components such as references to publications and molecular level structures (only the identities of those structures will be of interest when performing a comparison), and (2) replace document-specific identifiers of the items with their corresponding identifiers in their authoritative sources if they exist to ensure the uniqueness and consistency of the item identifiers.

B. The Similarity Scoring Mechanism

To indicate how similar a given pair of pathways is, a similarity score is generated from each comparison. The

comparison result of each pair of XML documents is used to calculate the similarity score. The similarity score is an indication of how similar/different two pathways are in terms of their compositions and structures/topologies. Our scoring mechanism is derived by extending the scoring function suggested by Koyuturk et al [40] assuming that the score is approximately normally distributed.

Suppose that M_s is the set of matched XML nodes (structure); M_c the set of matched objects (composition); N_s the set of mismatched nodes; and N_c the set of mismatched objects. Define the scoring function for matches as *m( )* being the sum of all matches, and that for mismatches as *n( )* being the sum of all mismatches. The similarity score *S* is calculated using the following

$$S = \sum_{p \in Ms} m(p) + \sum_{q \in Mc} m(q) - \sum_{r \in Ns} n(r) - \sum_{t \in Nc} n(t).$$

Note that this similarity scoring function does penalize mismatches, and therefore a similarity score may be negative, indicating the dissimilarity of the pathways being studied.

Matches and mismatches of pathway structures and components can be found in the DiffGrams, and a collection of similarity scores are derived from the comparisons of the pathway in question against all the known or chosen pathways in the pathway repository. The scores are ordered to rank the similarity of the pathway-pairs.

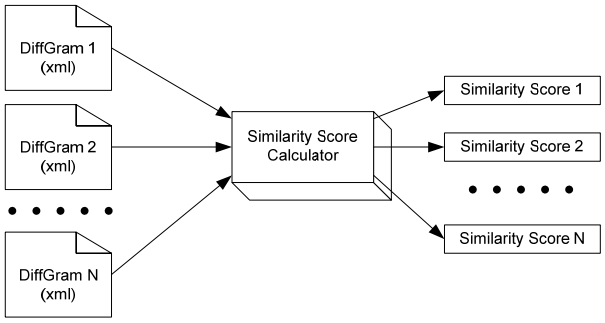

Fig. 2. Similarity scores are calculated from the results of comparing each pair of pathways. The results are in the format of DiffGram.

C. Implementation with High Performance Computing Resources

Our similarity search of biological pathways involves pair-wise comparison of hundreds or thousands pairs of XML documents. This can be a time-consuming process that requires a large amount of processor cycles. Pinter et al. reported that in a study of metabolic pathway similarity search between E. coli (113 pathways) and S. cerevisiae (151 pathways), it took 3.66 hours to perform the all-against-all alignments with an average of 47 seconds per query on a Pentium IV 2.6GHz machine with 512 MB of Memory [7]. In our case, it can take more time to perform the comparison given that we are dealing with several types of more sophisticated pathways and the XML presentation of the pathway data.

To overcome the need for intensive processor cycles, we leverage the existing Windows-based Condor [41] pools on our campus by distributing the task of comparing each pair of XML documents to individual Condor nodes. For a given pathway in question, a submit file is created to include a Condor "job cluster" that instructs the Condor submit host to copy to each node a pair XML documents and the xmldiff executable along with a dynamic-link library (DLL) of xmldiff, and directs the result of each comparison to be forwarded back to the submit machine where the similarity scores are calculated and rank-ordered. The requirement for Condor nodes capable of performing the comparison is any machine with Windows XP or higher installed to ensure the XML Diff and Patch can execute successfully in Microsoft .NET Framework 2.0.

III. DISCUSSIONS

The production version of our pathway similarity search system is still under development, the preliminary tested of the method is promising and feasible. The method seems to be relatively simple and easy to implement without loss of power and flexibility. A pathway in question can be queried against several different types of biological pathways; different criteria can be chosen to difference the XML-based pathway data; and different similarity scoring mechanisms can be readily applied to different types of similarity queries.

The performance and optimization of the pathway similarity search is being studied. Although there is no dependency for inter-processor communication between Condor nodes (which are CPUs residing in the computers in our Condor environment), the execution time of a each Condor "job cluster" is a function of (1) the time required to copy pairs of XML pathway datasets to an available Condor node, (2) the longest execution time of the pair-wise comparison among all the Condor nodes, (3) the time needed to forward the comparison results to the submit machine, and (4) the availability of idle Condor nodes. Transferring files over the network seems to be time-consuming and it can become worse if any Condo jobs are overflowed to the Condor pool at another campus due to the heavy usage of computers in the local Condor pool. Currently, all the pathway XML documents are preprocessed to eliminate the insignificant items which do not contribute to the calculation of the similarity score, reducing the sizes of the XML documents and thus improving the data transfer rate.

It is hoped that the production system of the pathway similarity search can aid biomedical scientists to address intriguing biomedical questions by performing pathway similarity searches within the same species or across different species. It is also hoped that our method can be applied to the other XML-format pathway datasets to encompass a more broad coverage of biological pathways, making possible the genomic scale analyses of biological

pathway datasets.

ACKNOWLEDGMENT

Authors wish to thank the Condor support team at the CS Department of University of Wisconsin, Madison for their timely help on many Condor technical issues.

REFERENCES

[1] Fishman, M.C. and J.A. Porter, Pharmaceuticals: a new grammar for drug discovery. Nature, 2005. 437(7058): p. 491-3.
[2] Bader, G.D., M.P. Cary, and C. Sander, Pathguide: a pathway resource list. Nucleic Acids Res, 2006. 34(Database issue): p. D504-6.
[3] Pathguide: the Pathway Resource List [http://pathguide.org].
[4] Altschul, S.F., et al., Basic local alignment search tool. J Mol Biol, 1990. 215(3): p. 403-10.
[5] Sharan, R. and T. Ideker, Modeling cellular machinery through biological network comparison. Nat Biotechnol, 2006. 24(4): p. 427-33.
[6] Kelley, B.P., et al., Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A, 2003. 100(20): p. 11394-9.
[7] Pinter, R.Y., et al., Alignment of metabolic pathways. Bioinformatics, 2005. 21(16): p. 3401-8.
[8] Bandyopadhyay, S., R. Sharan, and T. Ideker, Systematic identification of functional orthologs based on protein network comparison. Genome Res, 2006. 16(3): p. 428-35.
[9] Forst, C.V., et al., Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. BMC Bioinformatics, 2006. 7: p. 67.
[10] Zhang, Y., et al., Phylophenetic properties of metabolic pathway topologies as revealed by global analysis. BMC Bioinformatics, 2006. 7: p. 252.
[11] Kelley, B.P., et al., PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res, 2004. 32(Web Server issue): p. W83-8.
[12] Koyuturk, M., A. Grama, and W. Szpankowski, An efficient algorithm for detecting frequent subgraphs in biological networks. Bioinformatics, 2004. 20 Suppl 1: p. I200-I207.
[13] Koyuturk, M., et al., Pairwise alignment of protein interaction networks. J Comput Biol, 2006. 13(2): p. 182-99.
[14] Liang, Z., et al., NetAlign: a web-based tool for comparison of protein interaction networks. Bioinformatics, 2006. 22(17): p. 2175-7.
[15] Tun, K., et al., Metabolic pathways variability and sequence/networks comparisons. BMC Bioinformatics, 2006. 7: p. 24.
[16] Deville, Y., et al., An overview of data models for the analysis of biochemical pathways. Brief Bioinform, 2003. 4(3): p. 246-59.
[17] Schaefer, C.F., Pathway databases. Ann N Y Acad Sci, 2004. 1020: p. 77-91.
[18] Luciano, J.S., PAX of mind for pathway researchers. Drug Discov Today, 2005. 10(13): p. 937-42.
[19] Karp, P.D., et al., Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res, 2005. 33(19): p. 6083-9.
[20] The BioCyc Database. [http://biocyc.org].
[21] Kanehisa, M., The KEGG database. Novartis Found Symp, 2002. 247: p. 91-101; discussion 101-3, 119-28, 244-52.
[22] The KEGG Database. [http://www.genome.jp/kegg/].
[23] Vastrik, I., et al., Reactome: a knowledge base of biologic pathways and processes. Genome Biol, 2007. 8(3): p. R39.
[24] Joshi-Tope, G., et al., Reactome: a knowledgebase of biological pathways. Nucleic Acids Res, 2005. 33(Database issue): p. D428-32.
[25] The Reactome Database. [http://www.reactome.org].
[26] Jiang, K. and J. Robertson, A System for Exploring and Visualizing Biological Pathways from Large Scale Datasets, Proc. 28th IEEE EMBC (EMBC 2008). 2008, IEEE: Vancouver, Canada.
[27] Jiang, K., An Object Model Based Repository for Biological Pathways using XML Database Technology. International Conference of Computational Science (ICCS 2007), Lecture Notes in Computer Science (LNCS) 4488, Y. Shi Eds., © Springer-Verlag Berlin Heidelberg, pp.393-6, May 2007.
[28] W3C. W3C XML Query (XQuery). [http://www.w3.org/XML/Query/].
[29] Hermjakob, H., et al.: The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol. 22 (2004) 177-83
[30] Hucka, M., et al.: The systems biology markup language (SBML). Bioinformatics. 19 (2003) 524-31
[31] Lloyd, C.M., Halstead, M.D. and Nielsen, P.F.: CellML: its future, present and past. Prog Biophys Mol Biol, 85 (2004) 433-50
[32] Stromback, L., et al.: Representing, storing and accessing molecular interaction data: a review of models and tools. Brief Bioinform. 7 (2006) 331-8
[33] Wang, Y., D.J. DeWitt, and J.-Y. Cai, X-Diff: an effective change detection algorithm for XML documents, in 19th International Conference on Data Engineering. 2003, IEEE. p. 519 - 530.
[34] Al-Ekram, R., A. Adma, and O. Baysal, diffX: an algorithm to detect changes in multi-version XML documents, in 2005 conference of the Centre for Advanced Studies on Collaborative research 2005, IBM Press: Toranto, Ontario, Canada p. 1-11.
[35] Chen, Y., S. Madria, and S. Bhowmick (2004) DiffXML: Change Detection in XML Data Database Systems for Advanced Applications / Lecture Notes in Computer Science Volume 2973, 289-301
[36] Logilab Project xmldiff. [http://www.logilab.org/859].
[37] DecisionSoft xmldiff Tool. [http://software.decisionsoft.com/software/xmldiff.pl].
[38] IBM: XML Diff and Merge Tool.
[39] Microsoft: XML Diff and Patch GUI Tool.
[40] Koyutürk, M.G., Ananth; Szpankowski, Wojciech, Pairwise Local Alignment of Protein Interaction Networks Guided by Models of Evolution, in Research in Computational Molecular Biology. 2005, Springer Berlin / Heidelberg.
[41] Litzkow, M.J., Livny M., and Mutka M.W.: Condor-a hunter of idle workstations. Proc. 8th International Conference on Distributed Computing Systems, 1988, pp.104-111.