Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Sciences 176 (2006) 2771–2790
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

MDSM: Microarray database schema matching using the Hungarian method

Yi-Ping Phoebe Chen [a,b,*], Supawan Promparmote [a], Frederic Maire [c]

[a] *Faculty of Science and Technology, School of Information Technology, Deakin University, 221 Burwood Highway, Melbourne, Vic. 3125, Australia*
[b] *Australia Research Council, Centre in Bioinformatics, Australia*
[c] *Centre for Information Technology Innovation, Faculty of Information Technology, School of Software Engineering and Data Communications, Queensland University of Technology, Australia*

Received 7 April 2004; received in revised form 21 November 2005; accepted 29 November 2005

Abstract

Current microarray databases use different terminologies and structures and thereby limit the sharing of data and collating of results between laboratories. Consequently, an effective integrated microarray data model is required. One important process to develop such an integrated database is schema matching. In this paper, we propose an effective schema matching approach called MDSM, to syntactically and semantically map attributes of different microarray schemas. The contribution from this work will be used later to create microarray global schemas. Since microarray data is complex, we use microarray ontology to improve the measuring accuracy of the similarity between attributes. The similarity relations can be represented as weighted bipartite graphs. We determine the best schema matching by computing the optimal matching in a bipartite graph using the Hungarian optimisation method. Experimental results show that

* Corresponding author. Address: Faculty of Science and Technology, School of Information Technology, Deakin University, 221 Burwood Highway, Melbourne, Vic. 3125, Australia.
E-mail address: phoebe@deakin.edu.au (Y.-P.P. Chen).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2005.11.015

our schema matching approach is effective and flexible to use in different kinds of database models such as; database schema, XML schema, and web site map. Finally, a case study on an existing public microarray schema is carried out using the proposed method.

Keywords: Microarray database schema; Schema matching; Hungarian method; Similarity function

1. Introduction

Traditionally, molecular biology experiments were based on one gene at a time; this was a limitation in obtaining the total picture of a gene function. With the advent of DNA microarray technology, researchers are able to gain a better understanding of the interactions among thousands of genes simultaneously. Such technological innovation has led to new insights into fundamental biological problems such as; gene discovery, gene regulation, disease diagnosis, drug discovery, and toxicology [9–11,23,24].

However, a biological experiment, typically, requires tens or hundreds of microarray, where a single microarray generates between 100,000 and a million fragments of data [9–11]. The organisation of such a huge-volume of data, produced by microarray techniques, is one of the biggest challenges that scientists in bioinformatics are facing. Only a limited number of efficient and public databases are available to store microarray data (http://www.cbil.upenn.edu/RAD2, http://genex.sourceforge.net/, http://staffa.wi.mit.edu/chipdb/public/, http://www.ebi.ac.uk/arrayexpress/, http://genome-www5.stanford.edu/); however, existing public microarray databases have their own distinct storage structures and implementations, and different hardware platforms, DBMS, data models and data languages. In addition, these databases are created by different developers; unavoidably they might use different definitions and terms to describe the same domain or concept. Even though there are efforts to develop microarray data resources that correspond to the standard Microarray Gene Expression Data (MGED) ontology (http://www.cbil.upenn.edu/Ontology/MGED_ontology.html), their databases are still not in final shape. As a result, this hampers the sharing of data with other laboratories and the collating of experimental results. Fortunately, these limitations have been previously addressed in fields outside the life sciences, particularly in the realm of commercial business. One successful approach to elucidate these limitations is database integration.

An integrated microarray database has been proposed in our previous work [16]. One important task in our integrated architecture is to create global microarray schema. This can be done by taking schemas as input to produce