Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 41 (2008) 232–241
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

BioLattice: A framework for the biological interpretation of microarray gene expression data using concept lattice analysis

Jihun Kim [a,c], Hee-Joon Chung [a], Yong Jung [a], Kack-Kyun Kim [c], Ju Han Kim [a,b,*]

[a] Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110-799, Republic of Korea
[b] Human Genome Research Institute, Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110-799, Republic of Korea
[c] Department of Oral Microbiology, Seoul National University College of Dentistry, Seoul 110-799, Republic of Korea

Received 25 December 2006
Available online 1 November 2007

Abstract

Motivation. A challenge in microarray data analysis is to interpret observed changes in terms of biological properties and relationships. One powerful approach is to make associations of gene expression clusters with biomedical ontologies and/or biological pathways. However, this approach evaluates only one cluster at a time, returning long unordered lists of annotations for clusters without considering the overall context of the experiment under investigation.

Results. BioLattice is a mathematical framework based on concept lattice analysis for the biological interpretation of gene expression data. By considering gene expression clusters as objects and associated annotations as attributes and by using set inclusion relationships BioLattice orders them to create a lattice of concepts, providing an 'executive' summary of the experimental context. External knowledge resources such as Gene Ontology trees and pathway graphs can be added incrementally. We propose two quantitative structural analysis methods, 'prominent sub-lattice' and 'core–periphery' analyses, enabling systematic comparison of experimental concepts and contexts. BioLattice is implemented as a web-based utility using Scalable Vector Graphics for interactive visualization. We applied it to real microarray datasets with improved biological interpretations of the experimental contexts.

Keywords: DNA microarray; Gene expression; Clustering; Concept analysis; Concept lattice

1. Introduction

One of the challenges in DNA microarray data analysis is to extract biological meanings from massive amounts of gene expression data. Clustering has been one of the most successful methods for extracting coordinately regulated sets of genes [1,2]. The 'post-analytical challenge' of interpreting clusters using biological knowledge is under active investigation. Many Gene Ontology (GO)-based tools for gene expression analysis have been developed [3–9]. Several groups have proposed interpretation methods using biological pathways [10–13]. Gene Set Enrichment Analysis (GSEA) uses predefined gene sets and ranks of genes to identify significant biological changes in gene expression datasets [14,15].

Despite the undoubted importance of ontology and pathway-based annotation methods, they have limitations. The result, for example, is typically a long unordered list of annotations for tens or hundreds of clusters. The methods evaluate only one cluster at a time in a sequential manner without considering the informative association network of clusters and annotations. It is very time-consuming to read the massive annotation lists for a large number of clusters. Moreover, it is unthinkably hard to manually assemble the 'puzzle pieces' (i.e., the cluster–annotation

* Corresponding author. Address: Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110-799, Republic of Korea. Fax: +82 2 742 5947.
E-mail address: juhan@snu.ac.kr (J.H. Kim).
URL: http://www.snubi.org/software/biolattice/ (J.H. Kim).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.10.003

sets) into an 'executive summary' (i.e., the context of the whole experiment). Ideally, the assembly should involve eliminating redundant attributes and organizing the pieces in a well-defined order for better biological understanding and insight into the underlying 'context' of the experiment under investigation.

Here, we propose BioLattice, a mathematical framework based on concept lattice analysis to organize traditional clusters and associated annotations into a lattice of concepts for better biological interpretation of microarray gene expression data. Concept lattice analysis was introduced by Rudolf Wille [16]. The theoretical foundation rests on mathematical lattice theory. It studies how objects can be grouped hierarchically according to their common attributes.

BioLattice considers gene expression clusters as objects and annotations as attributes and provide a graphical summary of the order relations by arranging them on a concept lattice in an order based on set inclusion relation. By thinking in terms of concepts and contexts rather than in terms of individual clusters and annotations, this framework sets out the scope of conceptual clustering. The rest of this paper is organized as follows. In Sections 2.1–2.3, we introduce concept lattice theory in general and describe datasets, annotation methods and techniques for the construction of biological concept lattices. In Section 2.4, we propose two structural analysis methodologies that can be applied to a complex biological lattice to extract central and peripheral concepts and major sub-contexts of differing biological significance from the lattice. Section 3 describes the analysis results and how to read and navigate a biological lattice. Structural robustness of a lattice was evaluated. Finally, conclusions and future works are detailed in the last section.

2. Methods

2.1. Concept lattice

Context is a triplet (G, M, I) consisting of two sets G and M and a relation I between G and M. The elements of G are called the objects and the elements of M are called the attributes. To show that object g has attribute m, we write gIm or $(g,m) \in I$. For a set $A \subseteq G$ of objects, we define $A' := \{m \in M | gIm \text{ for all } g \in A\}$ (i.e., the set of attributes common to the objects in A). Correspondingly, for a set $B \subseteq M$ of attributes, we define $B' := \{g \in G |\ gIm \text{ for all } m \in B\}$ (i.e., the set of objects that have all attributes in B).

The concept analysis models concepts as units of thought, consisting of two parts. A concept of the context (G, M, I) is a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. We call A the extent and B the intent of concept (A, B). The extent consists of all objects belonging to the concept while the intent contains all attributes shared by the objects. The set of all concepts of the context (G, M, I) is denoted by $C(G, M, I)$. A concept lattice is drawn by ordering (A, B), which are defined as concepts of the context (G, M, I). The set of all concepts of a context together with the partial order $(A1, B1) \leqslant (A2, B2) :\leftrightarrow A1 \subseteq A2$ (which is equivalent to $B1 \supseteq B2$) is called a concept lattice.

We can regard A as defining gene expression clusters that share common knowledge attributes and B as defining the knowledge terms that are annotated to the clusters. The concepts are arranged in a hierarchical order so that the order of $C1 \leqslant C2 \leftrightarrow A1 \subseteq A2 \leftrightarrow B1 \supseteq B2$ is defined at $C1 = (A1, B1)$, $C2 = (A2, B2)$. Fig. 1 demonstrates a context (or a gene expression dataset) with clusters and annotations. Note that the relation matrix between objects (i.e., rows or clusters) and attributes (i.e., columns or annotations) can be represented by a directed graph (Fig. 1(b)) or a concept lattice with nonreduced (Fig. 1(c)) and reduced labeling (Fig. 1(d)). A concept lattice organizes all clusters and annotations of a relation matrix into a single unified structure with no redundancy and no loss of information. If $E1$ is a set of $\{(K2), (b,d,f,j)\}$ and $E2$ is a set of $\{(K1, K2), (b,f,j)\}$, then $E2$ subsides $E1$ because $\{K2\} \subseteq \{K1, K2\}$ and $\{b,d,f,j\} \supseteq \{b,f,j\}$ (Fig. 1(c)).

The top element of a lattice is a unit concept, representing a concept that contains all objects. The bottom element is a zero concept having no object. Specifically, the direct upper neighbors of the zero concept are called atoms and the direct lower neighbors of the unit concept are called coatoms. Fig. 1(c) and (d) are different visual representations of the same context (i.e., Fig. 1(a)). Fig. 1(d) demonstrates reduced labeling, where objects and attributes that can be omitted without losing information are omitted for easier reading. The extent of a concept is formed by collecting all objects that can be reached by descending line paths from the concept and vice versa to the intents. If a label of attribute A (object O) is attached to a certain concept, the attribute label occurs in all intent (extent) members of the concept, reachable by all descending (ascending) paths in the lattice from this concept to zero (unit) concept of the lattice.

In many applications, background knowledge may be available that can be used to model and analyze the data represented in a context [17]. Fig. 1(d)–(f) illustrates that background knowledge (or the GO trees in (e)) can be added easily to a concept lattice (d), returning an expanded concept lattice (f) (i.e., (d) + (e) = (f)).

2.2. Datasets

Four publicly available datasets were used to evaluate BioLattice. The mouse anti-GBM IgA nephropathy model (AGBM) dataset has 15 hybridizations at five time points with triplicates [18]. We used the 1112 genes showing significant temporal patterns by permutation analysis as described in the original manuscript. The human HeLa cell-division cycle (HCDC) dataset contains 26 hybridizations [19]. We used 2626 probes having pathway information. The yeast cell-division cycle (YCDC) dataset is a large collection of 59 time-course hybridizations, alpha factor, cdc15 and cdc28 [20]. We selected 2446 genes after removing all genes whose maximum minus minimum val-