Protein Classification from Protein-Domain and Gene-Ontology Annotation Information Using Formal Concept Analysis

Mi-Ryung Han[1], Hee-Joon Chung[1], Jihun Kim[1], Dong-Young Noh[2,3,*],
and Ju Han Kim[1,4,*]

[1] Seoul National University Biomedical Informatics (SNUBI),
[2] Department of Surgery,
[3] Cancer Research Institute,
[4] Human Genome Research Institute,
Seoul National University College of Medicine, Seoul, Korea
{gene0309,joonny96,djdoc,dynoh,juhan}@snu.ac.kr

Abstract. There are a number of different attributes to describe ontology of proteins such as protein structure, biomolecular interaction, cellular location, and protein domains which represent the basic evolutionary units that form protein. In this paper, we propose a mathematical approach, formal concept analysis (FCA), which toward abstracting from attribute-based object descriptions. Based on this theory, we present extended version of algorithm, tripartite lattice, to compute a concept lattice. By analyzing tripartite lattice, we attempt to extract proteins, which are related to domains and gene ontology (GO) terms from bottom nodes to the top of lattice. In summary, using tripartite lattices, we classified proteins from protein domain composition with their describing gene ontology (GO) terms.

1 Introduction

The theory of concept (or Galois) lattices (Wille, 1884) provides a natural and formal approach to discover and represent concept hierarchies (Carpineto et al., 1993). Conceptual data processing (also widely known as 'formal concept analysis') has become a standard technique in data and knowledge processing that has been applied for data visualization, data mining, information retrieval (using ontologies) and knowledge management. Concept (or Galois) lattice analysis represents patterns of intersection and inclusion among dual subsets of two sets of discrete elements (i.e. objects and attributes) (Mische et al., 2000).

Since concepts are necessary for expressing human knowledge, any knowledge management process benefits from a comprehensive formalization of concepts. Formal concept analysis (FCA) offers such a formalization by mathematizing the concept of 'concept' as a unit of thought constituted of two parts: extension and intension. If data are small, as compared with data bases in bioinformatics,

[*] Corresponding authors

Y. Shi et al. (Eds.): ICCS 2007, Part II, LNCS 4488, pp. 347–354, 2007.

formal concept data analysis shows how this abstract technique can unfold and better interpret the biological topics. Using formal concept data analysis, we can focus on exploratory data analysis with meaningful concept relationships. Therefore, in this paper, we approach protein classification using a new extension of lattice analysis - tripartite lattices - (Fararo et al., 1984; Mische et al., 2000) which is based on the formal concept analysis (FCA).

We use protein, protein domain and Gene Ontology (GO) (Ashburner et al., 2000) terms to show the intersections and inclusions among them by proposing tripartite lattices. Protein domains represent the basic evolutionary units that form protein. Multi-domain proteins can be made from single domain combination, and proteins with two or more domains constitute the majority of proteins in all organisms studied. Furthermore, domains that co-occur in proteins are more likely to display similar function or localization (Mott et al., 2002) than domains in separate proteins. Therefore we can classify similar protein functional groups from protein domain composition.

2 Methods

2.1 Formal Concept Analysis

We briefly introduce the basic notions of Formal Concept Analysis (Ganter et al., 1999).

Definition 1. *A formal context is a triple of sets (G, M, I), where G is called a set of objects, M is called a set of attributes, and $I \subseteq G \times M$. The inclusion (g, m) $\in$ I is read "object g has attribute m".*

Table 1. Example data set. Formal context of Proteins (G) and GO terms describing the proteins (M).

GO terms (M) Proteins (G)	DNA repair	Protein amino acid phosphorylation	Protein binding	ATPase activity
Protein 1	X	X		X
Protein 2	X		X	
Protein 3	X			X
Protein 4		X		
Protein 5		X	X	
Protein 6	X	X		X

→ In table 1, six proteins (G) are annotated with four GO terms (M) using gene ontology information: {(g, m): (Protein 1, DNA repair, Protein amino acid phosphorylation, ATPase activity), (Protein 2, DNA repair, Protein binding), (Protein 3, DNA repair, ATPase activity), (Protein 4, Protein amino acid phosphorylation), (Protein 5, Protein amino acid phosphorylation, Protein binding), (Protein 6, DNA repair, Protein amino acid phosphorylation, ATPase activity)}

Definition 2. *For $A \subseteq G$ and $B \subseteq M$: $A' = \{m \in M \mid \forall g \in A\ (gIm)\}$, $B' = \{g \in G \mid \forall m \in B\ (gIm)\}$.*

→ That is, in a formal context, duality relationship of a subset of proteins denoted by A, and a subset of GO terms denoted by B. Here, A′ is the set of GO terms common to the Proteins in A, and B′ is the set of Proteins which have all GO terms in B. For example, {Protein 1, Protein 3}′ = {DNA repair, ATPase activity}, {Protein 1}′ = {DNA repair, Protein amino acid phosphorylation, ATPase activity}, {DNA repair, ATPase activity}′ = {Protein 1, Protein 3, Protein 6} as shown in table 1.

Definition 3. *A formal concept of a formal context (G, M, I) is a pair (A, B), where $A \subseteq G$, $B \subseteq M$, $A' = B$, and $B' = A$. The set A is called the extent, and the set B is called the intent of the concept (A, B). The concepts of a given context are naturally ordered by the subconcept-superconcept relation defined by $(A1, B1) \leq (A2, B2)$: $\Leftrightarrow A1 \subseteq A2$ ($\Leftrightarrow B2 \subseteq B1$).*

→ In our case, we say (A1, B1) = {(Protein 1), (DNA repair, Protein amino acid phosphorylation, ATPase activity)}, (A2, B2) = {(Protein 1, Protein 3), (DNA repair, ATPase activity)} in table 1. Then subconcept-superconcept relation can be defined by the order of {Protein 1} ⊆ {Protein 1, Protein 3} and {DNA repair, ATPase activity} ⊆ {DNA repair, Protein amino acid phosphorylation, ATPase activity}.

We constructed a two-mode binary matrix according to the associations between six proteins and four gene ontology terms as shown in table 1 (entering 'X' into the matrix whenever a particular protein is annotated with a gene ontology term). Each protein has more than one explicit GO term. The basic lattice procedure applies two algebraic operations - intersection and inclusion - to a two-mode incidence matrix (Mische et al., 2000). First, all possible intersections between the rows of a two-mode matrix are calculated and generate all possible intersecting subsets of GO term mapping proteins. The complete set which the vector containing all 'X' is then added to complete the array of subsets, showing which subsets are included in larger subsets. This dual ordering of sets of proteins and GO terms constitute the lattice that can be visualized in a line diagram in which nodes representing subsets are linked to nodes representing the larger subsets in which they are included.

In this paper, all matrices were run through the BioLattice program designed by Jihun Kim. The output of BioLattice program is interpreted with four different color coded concepts; concept lattice is decomposed into four sub-structures based on core-periphery model. (Red: The core structure is defined to maximal sublattice according to size in which every element is upper bounds of an atom. Green: Except for the core, all lower bounds to each elements of core are communicating elements. Yellow: Independent (background) structure is defined to each sublattice that atom equals to coatoms. Gray: The other parts of concept lattice are defined to peripheral structure.).

G = { Protein 1, Protein 2, Protein 3, Protein 4, Protein 5, Protein 6 }
M = { DNA repair, Protein binding, Protein amino acid phosphorylation, ATPase activity }
I = { (Protein 1, DNA repair), (Protein 1, Protein amino acid phosphorylation), (Protein 1, ATPase activity), (Protein 2, DNA repair), (Protein 2, Protein binding)...}

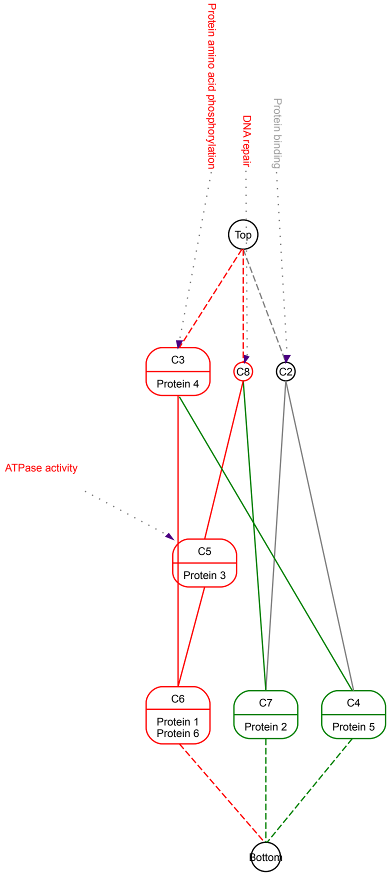

Fig. 1. Bipartite lattice: the concept lattice of the context in Table1 (Letter "C" stands for "concept")

Figure 1 presents a nine nodes lattice diagram based upon the protein by GO term matrix (6 X 4). This lattice diagram can be read in two directions, beginning at the top or the bottom. In here, {Protein 1, Protein 6} is a superconcept of {Protein 4} and {Protein 3} because {Protein 4} and {Protein 3} is described by a subset of the attributes describing {Protein 1, Protein 6}.

2.2 Tripartite Lattice: Interpenetrations Among Three Distinct Sets of Elements

Fararo et al. introduces tripartite structural analysis and shows how bipartite lattices can be extended to the tripartite lattices. They use persons, groups, and organizations; or persons, cultural systems, and social systems (Fararo et al., 1984). We investigated their analysis a step further to show the intersections and inclusions among three sets of interpenetrating biological elements (i.e. proteins, domains, GO terms). In this paper, tripartite lattices (theoretically generalizable to the k-partite level) show the interpenetration among three two-mode matrices: we use proteins by domains (PD), proteins by GO terms (PG), domains by GO terms (DG).

Let P, G and D denote the set of Proteins, GO terms and Domains respectively. We let the number of entities in the set X by nx (where X stands for P, G or D). Rxy denote the nx x ny-matrix in which Rxy $(i, j) = 1$ if the ith element of X is linked to the jth element of Y, and Rxy $(i, j) = 0$, otherwise. The matrix specifying the relationship between sets X and Y is the transpose of that representing the relationship between Y and X.
→ Rxy = Ryx′ (Rxx = 0, for each X)

Table 2. Matrix structure of a tripartite lattice

	GO	Protein	Domain
GO	0	R_{GP}	R_{GD}
Protein	R_{PG}	0	R_{PD}
Domain	R_{DG}	R_{DP}	0

Table 2 shows a symmetrical matrix with the upper right blocks composed of transposes of three lower left blocks. There are no within set relationships.

3 Results

3.1 Bipartite Lattice

Dataset. Exploratory data analysis is first performed using protein lists from Krebs_TCA_cycle pathway and Citrate_TCA_cycle pathway. We test bipartite lattice analysis using these pathways because they are one of the most investigated pathways. We use ArrayXPath which is publicly available major pathway resources including KEGG, GenMAPP, BioCarta and PharmGKB Pathways (Chung HJ et al., 2004). For further analysis, we have created a repository of protein, domain and gene ontology (GO) from SwissProt/TrEMBL for protein, InterPro for domain and Gene Ontology for GO term. In this paper, we use these resources to extract object and attribute information, and to perform formal concept analysis. The initial letter of protein ID is 'P' or 'Q' and Domain ID starts with three big letters 'IPR'.

Analysis of Bipartite lattice. We construct three two-mode binary matrices (see figure 2 (a), (b), (c)). These bipartite lattices show the relationship between protein and protein domain, protein and protein annotated GO term, domain and domain annotated GO term respectively (Each concept lattice of the matrices is not shown here). Through these lattices, we can point out which domains are common to several proteins or which GO terms are common to several proteins or which GO terms are common to several domains. However, the limitation of bipartite lattice analysis is that it presents only an abstract overview of the relations between two elements. For example, Krebs_TCA_cycle, Citrate_TCA_cycle pathways related protein 'P36957', 'Q02218' have the same GO term 'energy pathways'. If we want to know domain composition of those proteins, we have to search lattice twice with the same protein IDs (protein and protein domain, protein and protein annotated GO term). Then we can find that these proteins have different domains {P36957: IPR003016, IPR000069, IPR011053} {Q02218: IPR011603, IPR005475}. For further information, we have to search domain related GO terms in different concept lattice.

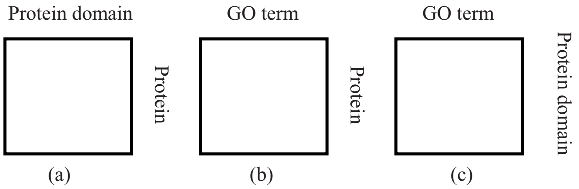

Fig. 2. Input matrix with attribute and object. (a) Two-mode binary matrix of Protein domain and Protein. (b) Two-mode binary matrix of GO term and Protein. (c) Two-mode binary matrix of GO term and Protein domain: Row elements are objects and column elements are attributes.

Therefore, bipartite lattice does not show us how these protein and domain elements come together with particular GO terms. If one more set of biological element is added to the bipartite lattice, which is called tripartite lattice, we can extract more compact and concrete information with three sets of biological elements (proteins, domains, GO terms). By proposing tripartite lattice, we can explore domain related proteins and their common GO terms simultaneously.

3.2 Tripartite Lattice

Dataset. We use Pathway crosstalk to select protein lists from random sampled pathways (see figure 3(a)) (Chung HJ et al., 2005). By random sampling, we choose a group of pathways from Pathway crosstalk with our fixed window size (see figure 3(b)). Random sampling approach is used to obtain more accurate estimates of data statistics. Then domains and GO terms are extracted using protein lists from below three random sampled pathways. We use distinct domains, proteins and GO terms to make matrix (see table 2). The GO terms are including domain annotated GO terms and protein annotated GO terms.

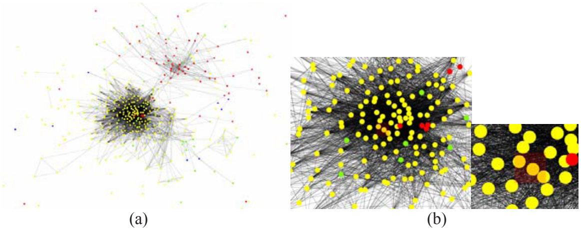

Fig. 3. Select specific pathways. (a) Pathway crosstalk: Calculating pairwise similarity matrix between each pair of pathways and applying multi-dimensional scaling method created the global crosstalk graph of major biological pathways. Yellow nodes represent BioCarta, green nodes GenMAPP, red nodes KEGG, and blue nodes PharmGKB Pathways. (b) Three pathways are selected by random sampling with our fixed window size. (Three neighboring pathways are chosen because we want to have as much common protein attributes as possible.) This is shown in transparent red rectangular region (BioCarta/Hs_IGF-1 Signaling Pathway, BioCarta/Hs_Insulin Signaling Pathway, BioCarta/Hs_Inhibition of Cellular Proliferation by Gleevec Pathway).

Analysis of Tripartite lattice. Formal concept analysis is performed using a symmetrical matrix with proteins, domains and GO terms. Among the most specific objects, all four proteins have domain "IPR000719" in common. Below lists are four proteins and their domain composition.
{P06213: IPR000719, IPR001245, IPR008266, IPR003961, IPR006212, IPR009030}
{P45983: IPR000719, IPR008351, IPR003527, IPR002290, IPR008271}
{Q13233: IPR000719, IPR008271, IPR002290, IPR007527}
{P27361: IPR000719, IPR008349, IPR003527, IPR002290, IPR008271}
We can find P45983 (Mitogen-activated protein kinase 8), Q13233 (Mitogen-activated protein kinase kinase kinase 1), P27361 (Mitogen-activated protein kinase 3) have the same synonym of EC 2.7.1.37 by searching ENZYME (Enzyme nomenclature database). Here, P06213 (Insulin receptor) has a synonym of EC 2.7.1.112 which is similar to EC 2.7.1.37 in that those enzymes can transfer a phosphate from a high energy phosphate such as ATP, to an organic molecule.
<Reaction catalysed>
EC 2.7.1.37: ATP + a protein ⇔ ADP + a phosphoprotein
EC 2.7.1.112: ATP + a protein tyrosine ⇔ ADP + a protein tyrosine phosphate

Therefore, we can classify above four proteins as a similar protein functional group. In addition, the function of each protein is described by GO terms in tripartite lattice. So we can say that domains that co-occur in proteins are more likely to display similar function or localization.

4 Discussion

We have investigated tripartite lattice, a new extension of lattice analysis, to show the interpenetrations among protein, domain and GO terms. By analyzing bipartite lattice, we extracted an abstract overview of the relations between two sets of biological elements. However, this analysis did not show us how protein and domain elements come together with particular GO terms. Using tripartite lattice, we classified proteins from protein domain composition with their describing GO terms. Because these proteins have similar functions, we can extract concrete information from tripartite lattice in that domains which co-occur in proteins are more likely to show similar function or localization as shown in GO term description.

By approaching extended version of algorithm to compute a concept lattice as we mentioned above, proteins are classified according to protein functions. However, the graphical representations of tripartite lattices are quite complex and difficult to read for large data sets. Therefore, we may consider interactive simplification of the obtained concepts by merging conceptual hierarchies as a future step.

Acknowledgement

This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (A060711), a grant from Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (01-PJ3-PG6-01GN07-0004), and a grant from Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (0412-MI01-0416-0002).

References

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25-29
2. Carpineto, C., Romano, G. (1993) GALOIS: An order-theoretic approach to conceptual clustering. Proceedings of 10th International Conference on Machine Learning, Amherst. pp. 33-40
3. Chung HJ, Kim M, Park CH, Kim J, Kim JH. (2004) ArrayXPath: mapping and visualizing microarray gene expression data with integrated biological pathway resources using Scalable Vector Graphics, Nucleic Acids Res. Jul 1;32:W460-W464.
4. Chung HJ, Park CH, Han MR, Lee S, Ohn JH, Kim J, Kim JH, Kim JH. (2005) ArrayXPath II: mapping and visualizing microarray gene expression data with biomedical ontologies and integrated pathway resources using Scalable Vector Graphics. Nucleic Acids Res.
5. Fararo, Thomas, J., Patrick Doreian. (1984) Tripartite structural analysis: Generalizing the Breiger-Wilson Formalism. Social Networks. 6, 141-175.
6. Ganter, B., Wille, R. (1999) Formal Concept Analysis: Mathematical Foundation. Springer, Heidelberg.
7. Mische, A., Pattison., P. (2000) Composing a civic arena: Publics, projects, and social settings. Poetics. 27, 163-194.
8. Mott, R., Schults, J., Bork, P., Ponting, C. P. (2002) Predicting protein cellular localization using a domain projection method. Genome Res. 12, 1168-1174.
9. Wille, R. (1884) Line diagrams of hierarchical concept systems. International Classification. 2, 77-86.