A User-Driven System for Navigating Protein Interaction Networks

J. H. Choi*, J. M. Park, J. Y. Jeong, S. H. Park
Electronics and Telecommunications Research Institute (ETRI)

1. Introduction

Proteins, as a cellular component in cell, may perform a few of molecular functions during biological processes. It may also have many interactions with other proteins to accomplish a biological process successfully in cell. In general, researcher employ the yeast two hybrid system for extracting the protein interactions, which is well known as a biological experiment [1][6].

Currently, as the experiment is performed in high throughput environments, many protein interactions of a few of species were generated. Therefore, they should be categorized according to their species or organism, and should be managed in database systematically. Since some of them are opened in Internet, we can access it by through World-Wide-Web. There are typical databases such as YPD (Yeast Proteome Database), PIMdb (Drosophila Protein Interaction Map database), BIND (Biological Interaction Network Database), DIP (Database of Interacting Protein), and etc.

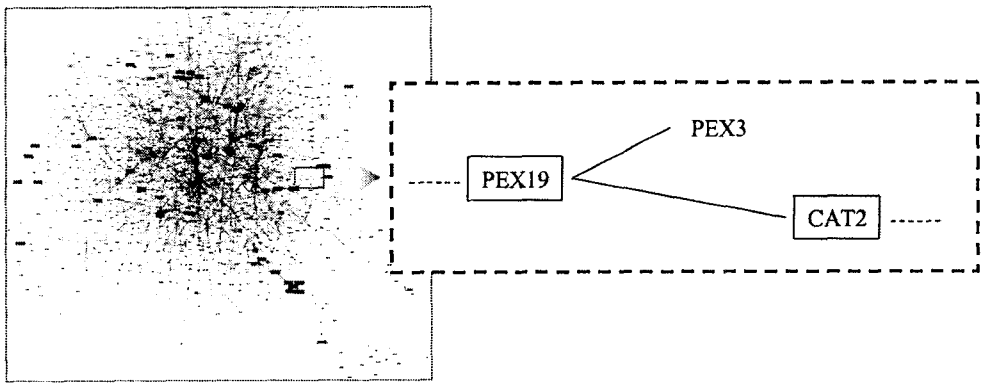

Fig. 1. Yeast Protein Network

In other hand, many interactions among proteins are defined as a network representing biological relationships in a cell. That is, proteins are represented as its vertices and interactions as its edges among vertices [3][4]. For example, [Fig. 1] depicts a relationship network among 2358 proteins existed in yeast cell [5]. The network takes vertices as yeast proteins PEX19, PEX3 and CAT2. They also have two edges among them. That is, protein PEX19 is interacted with both CAT2 and PEX3 simultaneous. Since a network representing all of interactions existed in a cell has the great number of vertices and edges, we can not analyze and interpret them without supporting the technique of computer visualizations.

In this paper, we propose a user-driven system for navigating protein interaction networks. Since it may drive users into specific information of network gradually, they can effectively explore it according to their intent. Our system consists of two core components; retrievals and visualization. The former performs a concept-based approach with Gene Ontology (GO), which can retrieve user-intended networks or proteins from database. The latter provides a multi-level approach for force-directed placements, which can automatically visualize the retrieved networks by optimized layout. For implementation of these functionalities, we employed an object-relationship database, ORACLE, which takes three object-relationship tables for network, protein and ontology respectively.

2. Concept-Based Network Retrievals

A network managed by database might fully or partially specify the interactions for a variety of species. It can make two networks to be different each other. In other word, two networks, describing a biological process, may be different because one seldom has the same proteins and interactions as others according to their species. Therefore, an interaction network can be identified by species name. Also it can be identified proteins and its biological characteristics. The biological characteristics of networks are classified as cellular component (CC), biological process (BP) and molecular function (MF). Since GO may hierarchically control the terms for specifying them, we employ it to retrieve networks by a concept-based method.

Ontology O is composed of three hierarchical relationships among standard terms which are categorized into BP(Biological Process), CC(Cellular Component) and MF(Molecular Function) respectively. Mostly these kinds of standard terms are adopted to specify the characteristics of proteins or networks. Therefore, user query is described by four facet queries, ORG, PRT, BP, CC and MF. ORG and PRT are simply represented by Boolean expressions for species names and protein names respectively. Others, BP, CC and MF, is used for ontology queries which are the Boolean expressions reformulated by linguistic terms in respective categories of O. In specific, a term in the queries is expanded as a disjunction of fuzzy terms, which are related with given term in the ontology.

A user query Q, a facet query, is evaluated as a fuzzy set |Q|. |Q| in a network of N is processed by the membership function $\mu_{|Q|}$: $N \rightarrow [0, 1]$. $\mu_{|Q|}(n)$ for each $n \in N$ denotes the degree of n to be member in the fuzzy set |Q|. Let A and B are two fuzzy subsets in N with membership functions μ_A and μ_B respectively. Standard min/max operation is used as operators between fuzzy sets. As a result, membership functions of $A \cup B$ and $A \cap B$ are formulated as $\mu_{A \cup B}(n) = \max(\mu_A(n), \mu_B(n))$ and $\mu_{A \cap B}(n) = \min(\mu_A(n), \mu_B(n))$ respectively.

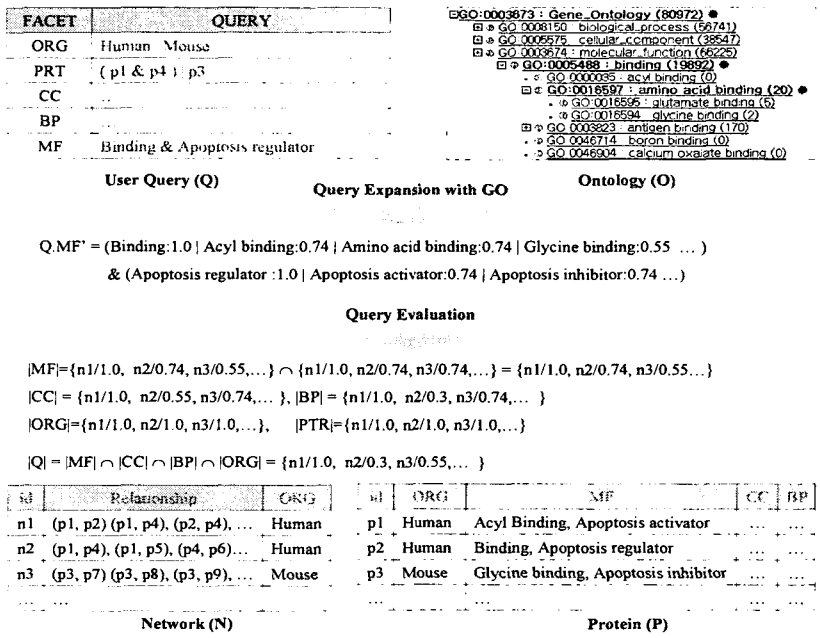

Fig. 2. Concept-Based Network Retrievals

[Fig. 2] is an example of our network retrieval processes using ontology. A user query Q includes five facet queries such as ORG="Mouse | Human", PRT="(p1 & p4) | p3" and MF="Binding & Apoptosis regulator". Since MF is a ontology query, our system may expand query terms according to MF hierarchical path in ontology. That is, a term 'Binding' in MF is translated into a disjunction of fuzzy terms, "Binding:1.0 | Acyl binding:0.74 | Amino acid binding:0.74 | Glycine binding:0.55". The translated terms take the fuzzy degree of relationships with original term 'Binding' in ontology. The fuzzy membership between two terms is evaluated by $\mu_{o_1}(o_2) = e^{-0.3 \times DIST(o_1, o_2)}$.

Since the species of n3 in network table N is 'Mouse' and p3, a protein, is involved in its relationship simultaneously, we can evaluate two facet query ORG and PRT as $\mu_{|ORG|}(n3) = \mu_{|PRT|}(n3) = 1.0$. In addition, MF characteristics of n3 include two ontology terms, 'Glycine Binding' and 'Apoptosis Inhibitor'. Therefore, n3 is evaluated as $\mu_{|MF|}(n3) = \min(0.55, 0.74) = 0.55$ because of $\mu_{|\text{'Glycine Binding'}|}(n3) = 0.55$ and $\mu_{|\text{'Apoptosis Inhibitor'}|}(n3) = 0.74$ in expanded MF query. In consequence, we can evaluate n3 for Q as $\mu_{|Q|}(n3) = \min(\mu_{|ORG|}(n3), \mu_{|PRT|}(n3), \mu_{|MF|}(n3), \mu_{|CC|}(n3), \mu_{|Q|}(BP)) = \min(1.0, 1.0, 0.55, 0.74, 0.74) = 0.55$. That is, n3 is related with user's intent by the degree 0.5.

3. Multi-Leveled Network Visualizations

For analysis and interpretation of the large sets of inter-related proteins in networks in detail, they should be visualized effectively as optimized layout. Most methods for visualizing protein networks are based on the FDP(Force-Directed Placement) algorithm[3]. It is divided into tree sub-modules; calculation of global forces, local forces and reposition. The calculation of global forces for a vertex evaluates a repulsive force against all vertices except its neighborhood vertices. The calculation of local forces for a vertex evaluates an attractive force for its neighborhood vertices. In subsequence, the reposition updates the position of a vertex to a minimal energy state, which is calculated by

reflecting two forces.

However, the FDP algorithm is not suitable to layout protein networks because they are consisted of many disconnected sub-networks. According to our experiment, the disconnections of networks prevent or slow down FDP from converging. In this paper, we make the advance of MFDP (Multilevel algorithm for Force-Directed Placement) [5]. It has two major processes; coarsening and expansion. Although there are many ways to create a coarser network, basically the coarsening process generates a simplified new network, which conserves the information of vertices and edges existed in the previous network. The process continued until the end conditions are satisfied. In the reverse direction, the expansion process recovers coarsened vertices and edges from its previous network and vertices are placed by using information of vertices which represented them in its previous network. At each level we use a FDP algorithm to layout each network. From the point of view of multi-level approach is attractive as it is an incrementally iterates to convergence which can reuse a previously calculated layout. The final network is visualized at user interface adequately.

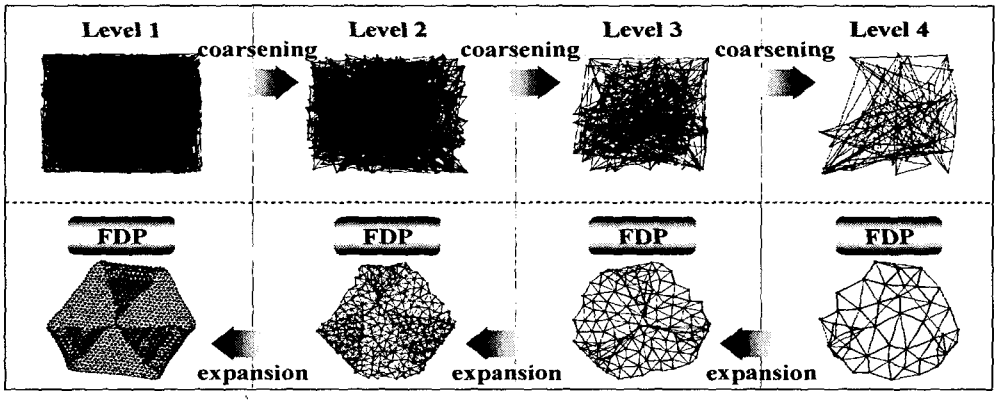

Fig. 3. process of multi-leveled network visualization

[Fig. 3.] shows process of multi-leveled network visualization. First, the process do coarsen network to define new network of next level and recursively iterate this procedure until the network size falls below some threshold. The coarsest network is then given. At each level each vertex has list of vertices that coarsened in previous level. Second, in the coarsest network, it runs FDP algorithm, which calculates good positions of vertex in this network. Third, it expands to the network of previous level. This time, calculated position in second process is used by vertices of expanded network. As applying this process repeatedly until initial network, we can get optimized layout of initial network. In this way, we can enhance the layout and decrease time to process of the complicated network.

4. Conclustions

The protein interactions are represented as a complex network which has many vertices and relationships among them. In the biological views, it is very valuable for users to explore it specifically. For this purpose, we proposed a user-driven system for navigating protein interaction networks, which may gradually drive users into specific information of a network according to their

intent. Our system is composed of two core components; retrieval and visualization. The former performs a concept-based approach with Gene Ontology (GO), which enables users to retrieve intended networks from database by a query. The latter provides a multi-level approach for force-directed placements, which can automatically visualize the retrieved networks as an optimized layout. In addition, it may facilitate the reference of protein information directly, which is realized by hyperlink included in a protein vertex.

References

1. C. L. Tucker, J. F. Gera, and P. Uetz, "Towards an Understanding of Complex Protein Interaction Maps," Trends in Cell Biology, Vol. 11, No. 23, 2001.
2. T. M. J. Fruchterman and E. M. Reingold, "Graph Drawing by Force-Directed Placement," Software: Practice and Experience, Vol. 21, No. 11, 1991.
3. P. Uetz, T. Ideker and B. Schwikowski, "Visualization and Integration of Protein-Protein Interactions," Cold Spring Harbor Laboratory Press, 2002.
4. S. Oliver, "Guilt-by-Association Goes Global," Nature-News and Views, Vol. 403, 2000.
5. C. Walshaw, "A Multilevel Algorithm for Force-Directed Graph Drawing," Graph Drawing 8th Intl. Symp, Berlin, 2001. [2] S. Field, and O. Song, "A Novel
6. S. Field, and O. Song, "A Novel Genetic System to Detect Protein-Protein Interactions," Nature 340: 245-247, 1989.