Poster G-13

A Wrapper-based Approach for Protein Identification in PPI Networks

Authors:
YongHo Lee (*Electronics and Telecommunications Research Institute (ETRI)*)
JaeHun Choi (*Electronics and Telecommunications Research Institute (ETRI)*)
MyungEun Lim (*Electronics and Telecommunications Research Institute (ETRI)*)
SooJun Park (*Electronics and Telecommunications Research Institute (ETRI)*)

Short Abstract: For the purpose of maintaining consistency of protein data among local databases, global databases, and PPI networks, we proposed a wrapper-based approach for protein identification in PPI networks. This method synchronizes a PPI network with a global database by identifying protein data of a global database using a wrapper.

Long Abstract:
PPI(Protein-Protein Interaction) networks are defined as the set of relationships among proteins. They are very important data which specify core items of one cell such as life cycle, DNA replication, transcription, signal transduction, and metabolism. The significance of PPI networks is getting increasing more and more because they are effectively used in the high value added bio-business field such as medical diagnosis or new drug discovery. These protein interaction data are extracted on a large scale by biological experiments using high-throughput technology and also frequently updated in global databases. Because of these updates, there can be inconsistent in protein data among global databases, PPI networks, and local databases constructed by themselves.
For resolving inconsistency of protein data, we propose a method which has two parts: one is the data consistency policy between a PPI network and a local database, and the other is synchronizing a PPI network with a global database by identifying protein data of a global database using a wrapper. As using the method, we can get consistency of a PPI network, a local database, and a global database. A PPI network always maintains up-to-date data because a wrapper provides real-time processing. Therefore, reliability of protein data is guaranteed. Our system is implemented as the following procedures. First, UniProt which has information to be extracted is selected. Second, a script file with interesting information extracted from UniProt is made. Last, the file is transformed into a wrapper. After that, the system executes the wrapper by using a wrapper API. Once the wrapper executes, it extracts interesting information from HTML pages in the form of text and transforms the extracted information of the text type into that of the XML type. In the recent biology field, XML-typed data are largely used in experiment results or data management for scalability and convenience. For that reason, the wrapper returns information in the form of XML. We constructed a local database by migrating Swiss-Prot. When a local database is updated by a global database periodically, it occasionally occurs that user-selected protein data from a PPI network are not consistent with those from a local database. For resolving this inconsistency, data from a PPI network must be synchronized with a local database. For example, the protein P1 of a PPI network is originally same as the protein P1 of a local database. Whenever P1 of a local database is updated as P1', there is inconsistency of the protein P1 between a PPI network and a local database. In this case, the proposed system

synchronizes P1 of a PPI network with P1' of a local database. When a global database is updated by the results of biological experiments, there can be inconsistent between a PPI network and a global database. Once inconsistency occurs, the system synchronizes both a local database and a PPI network with a global database by identifying protein data of a global database. We identified protein data using the UniProt wrapper and used UniProt as global database. UniProt is the union of Swiss-Prot as protein knowledgebase and TrEMBL with computer-annotated supplement to Swiss-Prot. The UniProt wrapper extracts protein information and has input data such as protein name, protein accession number, gene name, description, and organism. It extracts protein information such as protein name, protein accession number, last modified annotations, synonyms, gene information, TaxID, GO information, keywords, sequence, and so on. The sequence of identifying the protein is following. Specific proteins are selected from a PPI network, and then the synchronization component is executed. The system makes a query by using AND/OR operation of protein information which is inputted and then connects the UniProt SRS system. Once the query is executed in the UniProt SRS system, a protein accession number list is returned as a result. This list includes zero or more protein accession numbers. While the list has one or more protein accession numbers, the UniProt wrapper is executed repeatedly according to the numbers. After iterative execution, protein information of the UniProt is extracted in the form of text type and is transformed into the form of XML. The XML-type information is transformed into the string array list using DOM tree of the wrapper API. After that, the list is sent to a PPI network.

The proposed wrapper-based approach for protein identification makes possible global synchronization as well as local synchronization. Real-time synchronization guarantees the reliability of data by keeping protein data of a PPI network up-to-date. In the future, we will keep the research on the minimization of real-time synchronization time for protein identification and the exclusion of local synchronization by constructing a PPI network dynamically.