論文の内容の要旨
Dissertation Abstract

論文題目
Dissertation Title

Development of Interactive, Multi-Scale Network Navigation Method and Its Application to Functional Genomics Data

(対話型マルチスケールネットワークナビゲーション法の開発と機能ゲノミクスデータへの応

氏名
Name

プラーニーナラーラト タネート

In the post-genomic era, an exponential number of biological data are being produced at an accelerating pace by high-throughput technologies and avail- able via online databases on the Internet. Among these, binary relationship data that can be described as sets of elements and 1-to-1 associations (connections) between them have become increasingly common. Co-expressed gene pairs and protein-protein interactions exemplify this data type. Network (graph) visualization, where nodes and edges correspond to the elements and the connections respectively, is widely used for representing binary relationship data because it is expected to be more interpretable than a long list of associations. However, when network data are large and complicated (e.g., >100 edges), the network representations often become cluttered with jumbles of nodes and edges, known as "hair-balls", and thus fail to convey information effectively. Therefore, one of the key challenges is how to develop network navigation approaches that can abstract data properly and interactively, and visualize the data insightfully at a right level of detail. By such methods, researchers would be able to explore and interpret their large-scale networks much more effectively. Until recently, many studies have used various methods to tackle the cluttered-visualization problem, but still cannot obtain satisfactory results–truly clean and intuitive visualizations.

Hierarchical clustering is a technique that meaningfully and recursively groups data elements based on a similarity measure, thereby producing a hierarchy or tree of clusters. This method works with many types of data, including networks, to create groups of data elements in a multi-scale fashion. In the hierarchy, higher levels contain fewer, larger clusters with more data elements, or nodes in case of networks, than lower levels. Such a hierarchy can be applied to abstract the network visualization by showing only high-level clusters, thereby reducing the number of elements on the screen. By showing the actual members of each cluster at a certain level of the hierarchy, detailed information can be displayed at a particular scale. However, existing network visualization methods that offer such multi-scale navigation still have some drawbacks that hinder scientists from effectively and interactively exploring large biological network data, namely, (1) uses of clustering that depends upon user-provided information about hierarchies, (2) long running time (e.g., minutes to hours) required to abstract large networks, (3) inflexibility in navigation beyond fixed cluster boundaries, and (4) insufficiency of data abstraction, which leads to still cluttered network drawings.

In this dissertation, I present the *first* interactive, multi-scale navigation method for large and complicated biological networks and demonstrate its application to two types of functional genomics data, a yeast protein net- work dataset and an *Arabidopsis* gene co-expression dataset. The method is mainly composed of an ultrafast graph clustering technique that rapidly abstracts networks of about 100,000 nodes by recursively grouping densely connected portions and a biological-property-based clustering technique that uses property information provided for biological entities (e.g., Gene Ontology (GO) terms). It can rapidly and automatically abstract any region of large network data and produce biologically meaningful visualization with a manageable amount of information at all levels of detail. Apart from untangling large and complicated biological networks, it can be used to discover hidden knowledge in the networks readily and effectively as well. The method was first implemented as a stand-alone Java Swing application named NaviCluster (http://navicluster.cb.k.u-tokyo.ac.jp) and then integrated with Cytoscape as a plug-in, named NaviClusterCS (http://navicluster.cb.k.u-tokyo.ac.jp/cs/), to gain benefits from its usability and abundant useful features. I believe that the presented method will aid modern biologists in discovering knowledge from massive binary-relationship datasets more efficiently. In the final chapter, I anticipate the prospects for this research as four main directions: (i) clustering and implementation optimization, (ii) enhancement of functionalities, visualization, and user experiences, (iii) application to multiple types of networks, and (iv) integration with text mining toward interactive, systematic knowledge discovery.