Computational Molecular Biology and Genomics Project Proposal

Juchang Hua
David Koes
Zhenzhen Kou

Problem Statement

Because of the completion of the genome sequences of more and more organisms as well as the wide application of high-throughput techniques, the amount of protein interaction data is very large and is increasing exponentially.
To understand the behavior of these complicated networks, it is necessary to break them into small, easily understood, building blocks.

Network motifs, defined by Milo et al., have been accepted as a basic element of biological networks. Network motifs are simple topological patterns composed of a few nodes that appear many times in different places of the network. Motifs have been identified based on their higher likelihood of appearing a real biological network compared to a random network.

Project Goal

The goal of our project is to apply an algorithm from Berg and Lässig to protein-protein interaction networks derived from the 2-hybrid method. The advantage of this algorithm is that it can find motifs between similar topological patterns, not just identical patterns.

We want to find the motifs in a 2-hybrid network, compare the result with other procedures, and examine some motifs for biological meaning. If possible, we will improve the performance of the algorithm.

Project Plan

Data Source

We will use the yeast two-hybrid (Ito) protein interaction data linked off the course webpage. We'll first work on the Core data (841 interactions with more than 3 IST hits). This data will be converted into an intermediate graph representation used by our algorithm. Our proposed data representation is node-based. Each node consists of four attributes, each of which can represented by one line in an ASCII file:

```
node name, which is the gene name in the original Ito data
node information, which is the (Bait/Prey) description in Ito data
node number, which is a number assigned to the node uniquely
neighboring node numbers, space separated
```

If time permits, we will implement converters for other protein-protein interaction data sources such as yeast interaction data in DIP.

Computational Approach

We will implement the algorithm described in [Lassig03] and apply it to protein-protein interaction networks. This algorithm consists of three steps:

1. Find all the “interesting” sub-graphs of some fixed size *n*. In the original paper and “interesting” sub-graph were non-tree-like. Since our graphs are undirected,

we may not be able to prune the number of sub-graphs as much.
2. Compute the pair-wise minimal mismatch for all pairs of sub-graphs using a scoring function. The scoring function is the log likelihood of the alignment.
3. Use Monte-Carlo simulated annealing to find the best combination of pair-wise alignments yielding an approximation of the best multiple alignment.

The highest scoring multiple alignments of sub-graphs identify motifs. Furthermore, this approach identifies similar, yet not identical sub-graphs as belonging to the same motif.

Tasks to be performed

- Literature search (J)
- Conversion of data sets to intermediate form (J,Z)
- Implementation of algorithm including implementation of scoring functions
    - implement sub-graph finding routine (D)
    - implement simulated annealing using simplistic scoring functions (D)
    - add statistically significant scoring functions (Z)
- Evaluation of results (J)
- Based on evaluation improve the algorithm and repeat until we run out of time. (D,J,Z)

Division of work

Very generally, Juchang will deal with the biological aspects of the project, Zhenzhen with the statistical, and David with the computational programming. The individual tasks are allocated as labeled above.

Evaluation

Having run our algorithm on a data set, we will examine the motifs produced. We will show that the motifs found are indeed motifs (non-random sub-graphs that occur with high frequency in the original graph).

In order to be useful, the motifs found should represent some biologically meaningful unit. For example, they may identify a protein complex or a standard building block of a pathway. We will examine the found motifs to see if they match up with a known structure or if we can infer a new meaningful structure from them.

Deliverable

We will write up our methodology and results in an 8-12 page report and provide a link to a tarball of the relevant source code.

Lässig03 Lässig, Michael and Berg, Johannes *Local graph alignment and motif search in biological networks.* 2003