Network Analysis Tools (NeAT)
Tutorial

Sylvain Brohée
sbrohee@ulb.ac.be

Karoline Faust
kfaust@ulb.ac.be

Jacques van Helden
jvhelden@ulb.ac.be

Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
Laboratory of Genome and Network Biology
Université Libre de Bruxelles, Belgium
http://www.bigre.ulb.ac.be

December 10, 2007

Contents

Chapter 1

Introduction

Since a few years, large scale biological studies produced huge amounts of data about networks of molecular interactions (protein interactions, gene regulation, metabolic reactions, signal transduction). The integration of these data sets can be combined to acquire a global view of the pieces that, altogether, contribute to the complexity of biological processes. High-throughput data is however notoriously noisy and incomplete, and it is important to evaluate the quality of the different pieces of information that are taken in consideration for building higher views of biological networks.

An important effort will be required to extract reliable information from the ever-increasing ocean of high-throughput data. This will require the utilization of powerful tools that enable us to apply statistical analysis on large graphs. For this purpose, we developed the **Network Analysis Tools** (***NeAT***), as set of tools performing basic operations on networks and clusters.

The tools can be used in three ways:

1. **Web server interface**

   http://rsat.scmbb.ulb.ac.be/neat/

   The Web interface gives a convenient and intuitive access to the tools, and allows you to bring your data sets through some typical analysis work flows in order to extract the best of it.

2. **Stand-alone application**

   http://rsat.scmbb.ulb.ac.be/rsat/distrib/

   Most of the tools are freely available to academic users, according to a licence for non-commercial and non-military usage.

   The license covers both the Regulatory Sequence Analysis Tools (***RSAT***) and the Network Analysis Tools (***NeAT***). It can be downloaded from the RSAT Web site.

3. **Web services**

In addition, people having computer skills can also use be same tools via a Web services interface, in order to integrate them in automatic work-flows. To obtain information on the Web services, connect the ***NeAT***web server, and in the left menu, select **Information - Web services**.

Chapter 2

Comparisons between networks

2.1 Introduction

Protein interaction networks have deserved a special attention for molecular biologists, and several high-throughput methods have been developed during the last years, to reveal either pairwise interactions between proteins (two-hybrid technology) or protein complexes (methods relying on mass-spectrometry). The term *interactome* has been defined to denote the complete set of interactions between proteins of a given organism.

Interactome data is typically represented by an un-directed graph, where each node represents a polypeptide, and each edge an interaction between two polypeptides.

The yeast interactome was characterized by the two-hybrid method by two independent groups, Uetz and co-workers [**?**], and Ito and co-workers [**?**], respectively. Surprisingly, the two graphs resulting from these experiments showed a very small intersection.

In this tutorial, we will use the program ***compare-graphs*** to analyze the interactome graphs published by from Uetz and Ito, respectively.

We will first perform a detailed comparison, by merging the two graphs, and labelling each node according to the fact that it was found in Ito's network, in Uetz' network, or in both. We will then compute some statistics to estimate the significance of the intersection between the two interactome graphs.

2.2 Computing the intersection, union and differences between two graphs

2.2.1 Study case

In this demonstration, we will compare the networks resulting from the two first publications reporting a complete characterization of the yeast interactome, obtained using the two-hybrid method.The first network [**?**] contains 865 interactions between 926 proteins.The second network [**?**] contains 4,038 interactions between 2,937 proteins.We will merge the two networks (i.e. compute their union), and label each edge

according to the fact that it is found in Ito's network, Uetz' network, or both. We will also compute the statistical significance of the intersection between the two networks.

2.2.2 Protocol for the web server

1. In the ***NeAT***menu, select the command ***network comparison***.

   In the right panel, you should now see a form entitled "compare-graphs".

2. Click on the button `DEMO`.

   The form is now filled with two graphs, and the parameters have been set up to their appropriate value for the demonstration. At the top of the form, you can read some information about the goal of the demo, and the source of the data.

3. Click on the button `GO`.

   The computation should take a few seconds only. The result page shows you some statistics about the comparison (see interpretation below), and a link pointing to the full result file.

4. Click on the link to see the full result file.

2.2.3 Protocol for the command-line tools

If you have installed a stand-alone version of the NeAT distribution, you can use the program ***compare-graphs*** on the command-line. This requires to be familiar with the Unix shell interface. If you don't have the stand-alone tools, you can skip this section and read the next section (Interpretation of the results).

2.2.4 Interpretation of the result

The program ***compare-graphs*** uses symbols R and Q respectively, to denote the two graphs to be compared. Usually, R stands for reference, and Q for query.

In our case, R indicates Ito's network, whereas Q indicates Uetz' network. The two input graphs are considered equivalent, there is no reason to consider one of them as reference, but this does not really matter, because the statistics used for the comparison are symmetrical,as we will see below.

Union, intersection and differences

The result file contains the union graph, in tab-delimited format. This format is very convenient for inspecting the result, and for importing it into statistical packages (R, Excel, ...).

The rows starting with a semicolon (;) are comment lines. They provide you with some information (e.g. statistics about the intersection), but they will be ignored by graph-reading programs. The description of the result graph comes immediately after these comment lines.

Each row corresponds to one arc, and each column specifies one attribute of the arc.

1. **source**: the ID of the source node
2. **target**: the ID of the target node
3. **label**: the label of the arc. As labels, we selected the option "Weights on the query and reference". Since the input graphs were un-weighted, edge labels will be used instead of weights. The label `<NULL>` indicates that an edge is absent from one input network.
4. **color** and **status**: the status of the arc indicates whether it is found at the intersection, or in one graph only. A color code reflects this status, as indicated below.
   - *R.and.Q*: arcs found at the intersection between graphs R and Q. Default color: green.
   - *R.not.Q*: arcs found in graph R but not in graph Q. Default color: violet.
   - *Q.not.R*: arcs found in graph Q but not in graph R. Default color: red.

   The result file contains several thousands of arcs, and we will of course not inspect them by reading each row of this file. Instead, we can generate a drawing in order to obtain an intuitive perception of the graph.

Sizes of the union, intersection and differences

The beginning of the result file gives us some information about the size of the two input files, their union, intersection, and differences.

```
; Counts of nodes and arcs
;       Graph   Nodes    Arcs     Description
;       R       2937     4038     Reference graph
;       Q       926      865      Query graph
;       QvR     3215     4730     Union
;       Q^R     648      173      Intersection
;       Q!R     278      692      Query not reference
;       R!Q     2289     3865     Reference not query
```

Statistical significance of the intersection between two graphs

The next lines of the result file give some statistics about the intersection between the two graphs. These statistics are computed in terms of arcs.

```
; Significance of the number of arcs at the intersection
;      Symbol    Value     Description                         Formula
;      N         3215      Nodes in the union
;      M         5166505   Max number of arcs in the union     M = N*(N-1)/2
;      E(Q^R)    0.68      Expected arcs in the intersection   E(Q^R) = Q*R/M
;      Q^R       173       Observed arcs in the intersection
;      perc_Q    20.00     Percentage of query arcs            perc_Q = 100*Q^R/Q
;      perc_R    4.28      Percentage of reference arcs        perc_R = 100*Q^R/R
;      Jac_sim   0.0366    Jaccard coefficient of similarity   Jac_sim = Q^R/(QvR)
;      Pval      0         P-value of the intersection         Pval=P(X >= Q^R)
```

A first interesting point is the maximal number of arcs (M) that can be traced between any two nodes of the union graph. In our study case, the graph obtained by merging Ito's and Uetz' data contains $N = 4,730$ nodes. This graph is un-directed, and there are no self-loops. The maximal number of arcs is thus $M = N * (N-1)/2 = 5,166,505$. This number seems huge, compared to the number of arcs observed in either Uetz' ($N_Q = 865$) or Ito's ($N_R = 4,038$) graphs. This means that these two graphs are sparse: only a very small fraction of the node pairs are linked by an arc.

The next question is to evaluate the statistical significance of the intersection between the two graphs. For this, we can already computed the size that would be expected if we select two random sets of arcs of the same sizes as above ($N_Q = 865$, $N_R = 4,038$).

The probability for an arc to be selected in the first random set is $P(R) = N_R/M = 0.0007815728$. The probability for an arc to be selected in the second random set is $P(Q) = N_Q/M = 0.0001674246$. The probability for an arc to be drawn independently in both random sets is the product of the probabilities.

$$P(QR) = P(Q) * P(R) = 1.308545E^{-07}$$

The number of arcs expected by chance in the intersection is the probability multiplied by the maximal number of arcs.

$$E(QR) = P(QR)/M = Q/M \cdot R/M \cdot M = 1.308545E{-}07 \cdot 5,166,505 = 0.68$$

Thus, at the intersection between two random sets of interaction, we woudl expect on the average a bit less than one interaction. It seems thus clear that the 173 interactions found at the intersection between he two published experiments is much higher than the random expectation.

We can even go one step further, and compute the *P-value* of this intersection, i.e. the probability to select at least that many interactions by chance.

The probability to observe *exactly* x arcs at the intersection is given by the hypergeometrical distribution.

$$P(QR = x) = \frac{C_R^x C_{M-R}^{Q-x}}{C_M^Q} \tag{2.1}$$

where

- R is the number of arcs in the reference graph;
- Q i the number of arcs in the query graph;
- M is the maximal number of arcs;
- x is the number of arcs at the intersection between the two graphs.

By summing this formula, we obtain the P-value of the intersection, i.e. the probability to observe *at least* x arcs at the intersection.

$$Pval = P(QR >= x) = \sum_{i=x}^{min(Q,R)} P(X = i) = \sum_{i=x}^{min(Q,R)} \frac{C_R^i C_{M-R}^{Q-i}}{C_M^Q}$$

We can replace the symbols by the numbers of our study case.

$$\begin{aligned} Pval &= P(QR >= 173) \\ &= \sum_{i=x}^{min(865,4038)} \frac{C_{4038}^i C_{5166505-4038}^{865-i}}{C_{5166505}^{865}} \\ &\approx 0 \end{aligned}$$

The probability is not properly speaking zero, but it is smaller than the smallest probability that can be computed by our statistical library (this limit is $\approx 10E^{-321}$).

Summary

In summary, the comparison revealed that the number of arcs found in common between the two datasets (Ito and Uetz) is highly significant, despite the apparently small percentage of the respective graphs it represents (≈20% of Ito, and ≈4% of Uetz).

Chapter 3

Exercises

1. Using the tool the tool ***network randomization***, generate two random graphs of 1000 nodes and 1000 arcs each (you will need to store these random networks on your hard drive). Use the tool ***network comparison*** to compare the two random graphs. Discuss the result, including the following questions:

   (a) What is the size of the intersection ? Does it correspond to the expected value ?

   (b) Which P-value do you obtain ? How do you interpret this P-value ?

2. Randomize Ito's network with the tool ***network randomization***, and compare this randomized graph with Uetz' network. Discuss the result in the same way as for the previous exercise.