Homework 1 (chapters 3-4 of U. Alon "An introduction to systems Biology").

Date: 2014-04-30

Exercise 1

Divide the genes in YEASTRACT into regulated genes which do not directly regulate the transcription of any gene (themselves included) and transcription factors (all the other genes). Consider the regulated genes as output nodes, the transcription factors as internal nodes, and regulation as links.
How many output nodes, how many internal nodes, and how many directed links of different types between internal nodes and output nodes are there in this data set? Call the number of internal nodes N, the number of output nodes, N^{out}, the number of directed links between internal nodes M^* and the number of links from internal nodes to output nodes M^{out}.*

Answer:

	NODES	**LINKS**
Internal	$N^* = 307$	M^*= 9516
Output	N^{out}= 6418	M^{out}= 192456

(197 self-links) .. the data was downloaded on 18th of April 2014.

Exercise 2

(analytical) Consider a network with N nodes, where every directed link is present with probability p (independence assumes). What is the expected number of directed links expressed in N and p? Call this function E(p,N).

Answer:

Let $\boldsymbol{N^2}$ be the number of possible links throughout the network with **N** nodes (including self-links). Trying to find the expected number of directed links, we have $\boldsymbol{N^2}$ independent Bernoulli distributions, each with parameter **p** and we are looking for the expected total number of successes i.e. a binomial distribution with parameters $\boldsymbol{N^2}$ and **p** and the expected value is $\boldsymbol{p \cdot N^2}$**.**
Therefore: $\boldsymbol{E(p, N) = p \cdot N^2}$

Exercise 3

What is the value of p such that E(p,N) = M*? Call this value p*.*

Answer:

We know from the first question that $\boldsymbol{N^* = 307}$ and $\boldsymbol{M^* = 9516}$.
Therefore we have:

$$\boldsymbol{E(p^*, N^*) = M^* \rightarrow p^* \cdot N^{*2} = M^* \rightarrow p^* = \frac{M^*}{N^{*2}} = \frac{9516}{307^2} \approx 0.101}$$

Exercise 4

(analytical) In a network with N^* nodes, where every directed link is present with probability p^* what is the expected number of

(a) Auto-regulatory motifs?
(b) Feed-forward loop (FFL) motifs?
(c) Feed-back loop motifs?

Answer:

a) The average number of self-edges is equal to the number of edges E times the probability that an edge is a self-edge, which is 1/N. So, we have that:

$$\boldsymbol{< N_{self} > \sim E \cdot p_{self} \sim E/N}$$

We found that $E = p^* N^{*2} = 0.101 * 307^2 \approx 9519$
Therefore the number of self-edges (thus auto-regulatory motifs) is equal to:

$$\boldsymbol{< N_{self} >_{real} = \frac{9519}{307} \approx 31}$$

b) According to U. Alon on page 43 of the book, "…the average number of occurrences of a subgraph G in the network is approximately equal to the number of ways of choosing a set of n nodes out of N: about N^* for large networks, multiplied by the probability to get the 'g' edges in the appropriate places."

Subsequently we have:

$$\boldsymbol{< N_{FFL} >_{real} \approx a^{-1} N^{*3} p^{*3}},$$

where 'α' is a number that includes combinatorial factors related to the structure and symmetry of each subgraph and we know that α=1 for FFLs.

Finally we get,

$$\boldsymbol{< N_{FFL} >_{real} \approx 1 \cdot 0.0101^3 \cdot 307^3 \approx 29811}$$

c) The same is true for Feedback Loops, but this time we know that parameter α=3. Therefore we get:

$$< N_{FBL} >_{real} = \frac{N^{*3} p^{*3}}{3} \approx 9937$$

Exercise 5

Consider the YEASTRACT data of transcription factors only. What is the actual number of…

a) Auto-regulatory motifs?
b) Feed-forward loop (FFL) motifs?
c) Feed-back loop motifs?

Answer:

Python scripts are provided with names corresponding to each points' answer for this question. Here I show the numeral answers only:

a) Auto-regulatory motifs: 197
b) FFL motifs: 91208
c) FBLs: 11195

Exercise 6

(Partial conclusion) Discuss the discrepancies between the answers to points 4 and 5 above (if any).

Answer:

It is clear that the discrepancies between the answers from 4 and 5 above are based on the fact that auto-regulation and feed-forward loops are network motifs. This means that they appear much more in real networks than they do in random ones (see page 28 figure 3.1 for auto-regulation and page 44 table 4.1 for FFL and FBL, in E. coli network). On the other hand, the FBLs do not appear so much in real networks and that will become obvious after we calculate its Z-score, which will be significantly lower.
A good measure to show this would be to calculate the **Z-scores** of each pattern.

- For self-edges: $Z = \frac{<N_{self}>_{real} - <N_{self}>_{rand}}{\sqrt{<N_{self}>_{rand}}} = \frac{197-31}{\sqrt{31}} \approx 30$
- For FFLs: $Z = \frac{<N_{self}>_{real} - <N_{self}>_{rand}}{\sqrt{<N_{self}>_{rand}}} = \frac{91208-29811}{\sqrt{29811}} \approx 355$
- For FBLs: $Z = \frac{<N_{self}>_{real} - <N_{self}>_{rand}}{\sqrt{<N_{self}>_{rand}}} = \frac{11195-9937}{\sqrt{9937}} \approx 13$

The Z-scores prove my points.

Exercise 7

Redo points 4, 5 and 6 above for the FFL with the condition that node Z in the FFLs is an output node. Define your random graph null model for this situation and compute the relevant probabilities in terms of the counts $N^*, N^{out}, M^*, M^{out}$. Describe clearly what you do.

7.4 (considering Z as output)

FFL motifs:
For the Feed Forward loop motifs, considering that Z is an output node, we will have to rethink about the ways of picking the nodes and the links of the FFL.
Thus, we have $2 \cdot \binom{N^*}{2}$ ways of picking the internal nodes X and Y (2 times, since we can switch between their positions). And also, we have a probability p^* for the internal nodes and a probability p^{out} for the external ones. The former is known, about the latter we have that $p_{out} = \frac{M_{out}}{N^* \cdot N_{out}}$.
This stands because the maximum number of output links is $N^* \cdot N_{out}$, since each internal node can go to each output etc.
So, the expected number of links is $M_{out} = N^* \cdot N_{out} \cdot p_{out}$, which shows the value of p_{out}.

Based on the above we can now calculate the expected number of FFL motifs as follows:

$$< \boldsymbol{N_{FFL}} > \approx \boldsymbol{2 \cdot \binom{N^*}{2} \cdot N_{out} \cdot p^* \cdot p_{out}^2} \approx \boldsymbol{2 \cdot \frac{N^*(N^*-1)}{2} \cdot N_{out} \cdot \frac{M^*}{N^{*2}} \cdot \left(\frac{M_{out}}{N^* \cdot N_{out}}\right)^2} \approx \boldsymbol{600.000,}$$

7.5 (considering Z as output)

After running the scripts to calculate the FFLs, I see that the results is the following:
FFL motifs: 1842285

7.6 (partial discussion)

Again we can see about exercises “Redo 4” and “Redo 5”, that for the Feed-Forward loops, the numbers are quite higher in the real network than in a random one. That makes them network motifs.
Let’s calculate the Z-score for the FFLs only now:

- $Z = \frac{<N_{self}>_{real} - <N_{self}>_{rand}}{\sqrt{<N_{self}>_{rand}}} = \frac{1840000 - 600000}{\sqrt{600000}} \approx 1600$

Exercise 8

Based on the structure of the network described in point 1, explain why the null model as defined in point 4 is sub-optimal for estimating the number of feedback loops. Propose a better graph null model and redo points 4 and 6 for the feedback loop.

Answer:

The null model that was described in point 4 does not take into account the topology of the network and it would be a good idea for our new null model to include both the output nodes as well as the output links.
Therefore, I will propose that we need to consider the whole network.
This means that there will also be output nodes (namely, nodes with no children). Therefore, we now have a number of N total nodes ($N^* + N^{out}$) and a number of M total links ($M^* + M^{out}$). Thus, N = 6725 and M = 201972.
The expected number of edges is $E(p, N) = p \cdot N^2$,
so we get that:

$$p \cdot N^2 = M \rightarrow \boldsymbol{p \approx \frac{M}{N^2} \approx \frac{201972}{6725^2} \approx 0.004}$$

8.4

- *Auto-regulatory motifs :* $< N_{self} > \sim E \cdot p_{self} \sim E/N \sim \frac{201972}{6725} \approx 31$
- *FFL motifs :* $< N_{FFL} > \approx a^{-1} N^n p^g \approx 1 \cdot N^3 \cdot p^3 \approx 6725^3 \cdot 0.004^3 \approx 19465$
- **FBLs** : $\boldsymbol{< N_{FBL} > \approx a^{-1} N^n p^g \approx 6488}$

8.6

Once again, we can see that auto-regulation and feed-forward loops, appear much more often on real networks than they do on random ones, something that is not true for FBLs. Let's calculate the Z-scores again.

- **FBLs:** $\boldsymbol{Z = \frac{<N_{self}>_{real} - <N_{self}>_{rand}}{\sqrt{<N_{self}>_{rand}}} \approx \frac{11195 - 6488}{\sqrt{6488}} \approx 58}$

Exercise 9

Redo point 4 for the three 4-node generalizations of the feed-forward loop defined in Alon Fig 5.6b page 83. Describe clearly what you do. (Optional) Redo points and 6 for those three cases.

Answer:

The 4-node generalizations of the feed-forward loops defined in the book (Fig 5.6b pg.83), all consist of 4 nodes and 5 links.

Now, considering that in exercise 4 we only take into account internal nodes, the symbols N* and p* will be used once again to denote that fact. The numbers corresponding to N*/p* have been calculated in previous exercises...

The number of permutations for the various FFLs is the same for all of the three cases, as we will see soon.

It is clear that we have 4 nodes in total and 2 nodes that are on the same level (are doing the same job), which means that we have 2 nodes that can be selected from 4, namely $\binom{4}{2}$.
Due to the symmetry of the resulting subgraphs this quantity should be multiplied by 2. This can be understood better if we consider that all these graphs are vertically symmetrical i.e if you put a vertical mirror in the middle of each one, and each side will be the same. Therefore the permutations that we can get for each graph are: $2 \cdot \binom{4}{2} = 2 \cdot \frac{4!}{2! \cdot (4-2)!} = 2 \cdot 6 = 12.$

Finally we get that:

$$< N >= 12 \cdot \binom{N^*}{4} \cdot {p^*}^5 \approx 45773$$

If I were to use Alon's formula just like in the book, I would have that α=2:

$$< N >= 2^{-1} \cdot {N^*}^4 \cdot {p^*}^5 \approx 46680$$

this is because we calculate the same subgraphs 2 times for each case of the 4-node generalizations.

The importance of the "symmetry factor".

For the FeedBack loops, we had that α=3, because the way that Alon's formula calculates the mean number of subgraphs for the FBLs, we would count the same FBL three times if not for the symmetry factor. It is also a convenient way to avoid the mathematically complex $\binom{n}{k}$ formula.
In the case that we use this combinatorial formula however, this symmetry factor is included in the calculations, just like I demonstrated in exercise 7.4.