Global Topological Study of Protein-protein Interaction Networks

Ka-Lok Ng (呉家樂)
Institute of Bioinformatics
Taichung Healthcare and Management University
Content

1. Introduction - Biological Networks
2. Protein-protein interaction database, DIP
3. Methodology
   – random graph, scale-free network
   – topological parameters + Shannon entropy
4. Results
5. Conclusions and Discussions
1. Biological Networks

Genomes of more than 30 species and at least 100 microbes have been completely sequenced. By now, scientists have genomic and a lot of proteins data. http://www.er.doe.gov/production/ober/microbial.html

Types of biochemical networks:
(i) gene regulatory network
(ii) protein-protein interaction network, and
(iii) metabolic network.
1. Biological Networks

Protein-protein interaction network

Proteins perform distinct and well-defined functions, but little is known about how interactions among them are structured at the cellular level. Protein-protein interaction network, such as binding interactions and formation of protein complex.

Limitation:
No subcellular location, and temporal information.

www.utoronto.ca/boonelab/proteomics.htm
1. Biological Networks

**Protein-protein interaction network**
- Experiment – **Yeast two-hybrid method**
- **protein-protein interactions are not random**
- highly connected proteins are *unlikely to interact* with each other.

The network of 318 interactions among the 329 proteins that are present in the nucleus of yeast cells. Proteins are represented by dots and their interactions by lines.

Reference: Maslov S, and Sneppen K., Science, 296, pp. 910-913, 2002.
2. Input data – Database of Interacting Protein (DIP)

DIP  [http://dip.doe-mbi.ucla.edu](http://dip.doe-mbi.ucla.edu)

DIP is a database that documents experimentally determined protein-protein interactions. We analyze the latest version, Feb., 2004, for seven different species, *S. cerevisiae*, *E. coli*, *H. pylori*, *C. elegans*, *D. melanogaster*, *H. sapiens* and *M. musculus*.
2. Input data – Database of Interacting Protein (DIP)

| Organism               | Proteins | Interactions |
|------------------------|----------|--------------|
| *S. cerevisiae* (CORE) | 2631     | 6558         |
| *H. pylori*             | 710      | 1420         |
| *E. coli*               | 429      | 516          |
| *C. elegans*            | 2638     | 4030         |
| *H. sapiens*            | 946      | 1129         |
| *M. musculus*           | 324      | 283          |
| *D. melanogaster*       | 7066     | 21017        |

- minimize experimental uncertainty ➔ CORE subset of budding yeast, *S. cerevisiae*
- according to the criteria described in Deane CM, Salwinski L, Xenarios I, Eisenberg D. (2002). Protein Interactions. *Mol Cell Prot* 1, 349.
3. Methodology – Graph Theory

Network representation. A network (graph) consists of a set of elements (vertices) and a set of binary relations (edges).
3. Methodology – Random Graph
= Graph Theory + Probability

Random graph (Erdos and Renyi)

Connectivity distribution $P(k)$

**Random network**

$P(k)$ vs. $k$ is a Poisson distribution

**Scale-free network**

$P(k) \sim k^{-\gamma}$ power-law distribution

Growth model of scale-free networks
- Addition of new nodes
- Preferential attachment

http://physicsweb.org/box/world/
3. Methodology

\[
M = \begin{pmatrix}
0 & 1 & \infty & \infty \\
1 & 0 & 1 & 1 \\
\infty & 1 & 0 & \infty \\
\infty & 1 & \infty & 0
\end{pmatrix}
\]

- \(\infty\) means not directly connected
- node \(i\) connectivity, \(k_i = \text{count}_j(m_{ij} = 1)\)

The shortest path
- Floyd algorithm, an \(O(N^3)\) algorithm. For iteration \(n\),
  \[M_{ij}^n = \min\{M_{ij}^{n-1}, M_{ik}^{n-1} + M_{kj}^{n-1}\}\]

- search for all possible paths,
e.g. 1-2, 1-2-3, 1-2-4, 2-3, 2-4
3. Methodology – Shannon Entropy

Shannon entropy, $H$
- Probability of observing a particular symbol or event, $p_i$, with in a given sequence
- Consider a binary system, an element $X$ has two states, 0 or 1

$$H = - \sum p_i \log p_i$$

- $H$ measure the “uncertainty” of a probability distribution

References:
1. [http://www.cs.unm.edu/~patrik/networks/PSB99/genetutorial.pdf](http://www.cs.unm.edu/~patrik/networks/PSB99/genetutorial.pdf)
2. [http://www.smi.stanford.edu/projects/helix/psb98/liang.pdf](http://www.smi.stanford.edu/projects/helix/psb98/liang.pdf)
3. [plus.maths.org/issue23/ features/data/](http://plus.maths.org/issue23/ features/data/)
Relative Entropy

\[ H_R = - \sum_i p_i \log_2 p_i \]

Fig. 3 Shannon entropies for a 2-state information source. Since the sum of the state probabilities must be unity, \( p(1) = 1 - p(0) \) for 2 states.
3. Methodology

Interaction path length or average diameter, $d$

- the average of the distances between all pairs of nodes
- determined by using the Floyd’s algorithm, the diameter $d$ is given by

$$d = \frac{\sum_j jf(j)}{\sum_j f(j)}$$

- where $j$ is the shortest path length
- frequency of the shortest interaction path length, $f(j)$
4. Results

Table 2. The average connectivity $<k>$, maximum connectivity $k_{max}$, average diameter $d$, average random diameter $<d_{rand}>$, average perturbed diameter $<d_{pert}>$ and the $\Delta$ results for the seven species.

| Species          | $<k>$ | $k_{max}$ | $d$  | $<d_{rand}>$ | $<d_{pert}>$ (\(\Delta\)) |
|------------------|------|-----------|------|--------------|-----------------------------|
| S. cerevisiae(CORE) | 4.87 | 111       | 5.01 | 4.01         | 5.01 (0.0%)                |
| H. pylori        | 3.87 | 54        | 4.14 | 3.72         | 4.14 (0.0%)                |
| E. coli          | 3.02 | 54        | 3.22 | 5.94         | 3.40 (5.6%)                |
| C. elegans       | 1.91 | 187       | 4.81 | 4.43         | 4.81(0.0%)                 |
| H. sapiens       | 2.30 | 33        | 6.05 | 5.61         | 6.16 (1.8%)                |
| M. musculus      | 1.67 | 12        | 3.58 | 6.64         | 3.74 (4.7%)                |
| D. melanogaster  | 5.90 | 80        | 4.46 | 5.20         | 4.46(0.0%)                 |
4. Results

Fig. 1. The logarithm of normalized frequency of connectivity vs the logarithm of connectivity for E. coli, H. pylori and S. cerevisiae.

Fig. 2. The logarithm of normalized frequency of connectivity vs the logarithm of connectivity for H. sapiens, M. musculus and D. melanogaster.
4. Results

Fig. 3. The relative Shannon entropy for single cellular organisms (S. cerevisiae, S. cerevisiae (CORE), H. pylori and E. coli) and the multi-cellular organisms (C. elegans, H. sapiens, M. musculus and D. melanogaster).

Table 3. The regression coefficient $\gamma$, Pearson coefficient $r$, coefficient of determination $r^2$, total degree of freedom $n$, and the relative Shannon entropy $H_R$ for the DIP data.

| Species             | $\gamma$ | $r$  | $r^2$ | Total dof | $n$  | $H_R$     |
|---------------------|----------|------|-------|-----------|------|-----------|
| S. cerevisiae(CORE) | 2.0±0.1  | 0.95 | 0.91  | 44        | 60   | 0.601(0.635) |
| H. pylori           | 1.7±0.1  | 0.95 | 0.90  | 30        | 40   | 0.606     |
| E. coli             | 1.5±0.4  | 0.84 | 0.70  | 9         | 12   | 0.541     |
| C. elegans          | 1.6±0.1  | 0.92 | 0.84  | 49        | 50   | 0.405     |
| H. sapiens          | 2.1±0.1  | 0.96 | 0.93  | 19        | 20   | 0.513     |
| M. musculus         | 2.4±0.2  | 0.97 | 0.93  | 10        | 10   | 0.314     |
| D. melanogaster     | 1.90±0.02| 0.96 | 0.93  | 76        | 78   | 0.577     |
4. Results

These highly connected proteins are pair-wise compared in an all-against-all manner using gapped BLAST (16), and none of the sequences shown significant sequences similarity (E-value < 0.001) except the tryptophan protein and SEC27 protein, nuclear pore protein, 26S proteasome regulatory particle chain and DNA-directed RNA polymerase.
5. Conclusions and Discussions

1. The $\log[P(k)]$ vs $\log[k]$ study indicates that protein-protein interactions form a scale-free network ($\gamma$ lies between 1.5 and 2.4)

2. We adapted the Shannon entropy approach to quantify the connectivity distribution, the result seems to suggest that multi-cellular organisms tend to have a lower relative entropy value for the single cellular organisms except for $D. melanogaster$. 
5. Conclusions and Discussions

Limitation
• Take the *spatial (subcellular localization)* and *temporal effects* into consideration, that is, not all the proteins will be located at the same organelle at the same time.
• Not able to detect *weakly interacted proteins*.

Discussions
• consider *multi-body interactions* in the protein network and find *clusters of proteins* that have many interactions among themselves. Such clusters correspond to *protein complexes*.
• *if two proteins share significantly large number of common partners* than random, they *could have close functional associations*.
• Structural study of *proteins have a highly connectivity* value
• Protein-protein interaction ➔ domain-domain interaction
Collaborator

Dr. Chien-Hung Huang (黃建宏)
Department of Information Management
Ling Tung College
Thank you.
Yeast Two-hybrid System

Yeast two-hybrid method identifies interacting proteins

http://cmbi.bjmu.edu.cn/
References

- Bose I. 2002. arxiv:cond-mat/0202192.
- Erdos P and Renyi A. 1960. Publ. Math. Inst. Hung. Acad. Sci. 5, 17.
- Farkas I., Jeong H., Viscek T., Barabasi A.L and Oltvai Z.N., 2003. Physica A, 318 601. (arxiv:cond-mat/0205181)
- Floyd R. W. 1962. Algorithm 97: Shortest path. Communication of the ACM 5, no.6, 345.
- Hasty J, McMillen D, Isaacs F. and Collins J. 2001. Nature Reviews Genetics 2, 268.
- Ideker T., Galitski T., and Hood L. 2001. Annu. Rev. Genomics Hum. Genet., 2, 343.
- Jeong H., Tombor B., Albert R., Oltvai Z.N., and Barabasi A.L., 2002. Science 295, 1662.
- Kanehsa M. 2000. Post-genome Informatics. Oxford University Press, Oxford.
- Kauffman, S.A. 1979. Kinetic Logic : A Boolean Approach to the Analysis of Complex Regulatory Systems, vol. 29 of Lecture Notes in Biomathematics, 30–60, Springer-Verlag, Berlin.
- Kauffman, S.A. 1993. The Origins of Order. Oxford University Press, Oxford.
- Krapivsky P.L., and Redner S. 2001. Phys. Rev. E, 63, 066123.
- Lewin B. 2000. Genes VII. Oxford University Press, Oxford.
- Lodish H., Berk A., Zipursky S., Matsudaira P., Baltimore D. and Darnell J. 2001. Molecular Cell Biology, 4th ed., W.H. Freeman, N.Y.
- Ma Hongwu and Zeng An-Ping, 2003. Bioinformatics, 19, 270.
- Maslov S, and Sneppen K. 2002. Science, 296, 910-913.
- Melian C. and Bascompte J., 2002. Ecology Letters 5, 705.
- McAdams H. and Arkin A. 1999. Trends in Genetics 15, 65.
- Tohsato Y., Matsuda H., and Hashimoto, A. in International Conference on Intelligent Systems for Molecular Biology (ISMB). AAAI Press, 2000.
- Wagner A. 2001. Mol. Biol. Evol. 18(7), 1283.
- Xenarios I., Frenandez E., Salwinski L., Duan X., Thompson M., Marcotte E., and Eisenberg D., 2001. Nucl. Acid Res. 29, 239.