

The spectral analysis on biology networks

Jiao Gu

Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig

Bled, February 16, 2012

Outline

1 Background

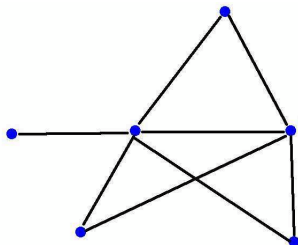
Outline

- 1 Background
- 2 Biology networks and spectrum metric

Motivation

We want to define a good metric to measure the difference between graphs, without identical information for each node.

Graph



Definition

A graph is an ordered pair $G = (V, E)$ comprising a set V of vertices, together with a set E of edges, which are 2-element subsets of V .

The spectrum of normalized laplacian matrix

Definition

The normalized Laplacian matrix $\Delta = [a_{ij}]$ has the form

$$a_{ij} = \begin{cases} 1, & \text{if } i = j \text{ and } n_i \neq 0, \\ -\frac{1}{n_j}, & \text{if } i \sim j \text{ is an edge,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The spectrum of normalized laplacian matrix

Definition

The normalized Laplacian matrix $\Delta = [a_{ij}]$ has the form

$$a_{ij} = \begin{cases} 1, & \text{if } i = j \text{ and } n_i \neq 0, \\ -\frac{1}{n_j}, & \text{if } i \sim j \text{ is an edge,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Definition

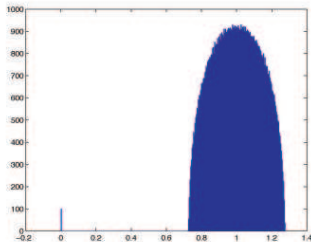
We call λ an eigenvalue of Δ if there exists some $u \neq 0$, such that

$$\Delta u = \lambda u$$

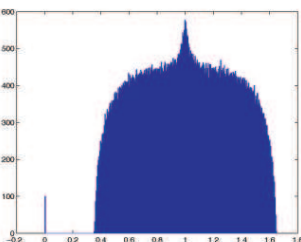
Normalized Laplacian matrix

- All the eigenvalues are bounded in $[0, 2]$, which makes it convenience to compare different networks.
- The normalized Laplacian matrix has been studied by several authors from the view of geometric analysis. It is an analogous of the Laplace-Beltrami operator in Riemannian geometry. As in Riemannian case, the eigenvalues λ here also represent important properties of the underlying graph, such as, connectivity, bipartition.

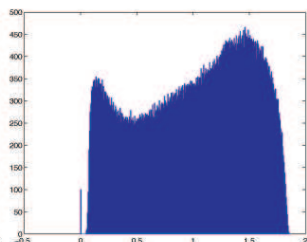
Spectrum of normalized Laplacian matrix



random network



scale-free network



small world network

Different types of networks' spectrum by *A. Banerjee, J. Jost 2008*

Eigenvalues sets

Homo sapiens (5923): $\{0^{\{632\}}, 0.003413, 0.0058298, 0.006595, 0.010495, 0.015472, 0.015935, 0.016366, 0.019801, 0.020154, 0.020775, \dots, 0.95121, 0.95259, 0.95513, 0.96187, 0.96311, 0.96845, 0.9724, 0.97805, 0.98627, 0.99116, 1^{\{2235\}}, 1.0528, 1.0569, 1.0596, 1.0614, 1.0621, 1.0634, 1.0661, 1.0692, 1.0698, 1.0782, \dots, 1.9723, 1.9767, 1.9789, 1.9792, 1.9802, 1.984, 1.9842, 1.9843, 1.9932, 1.9935, 2^{\{567\}}\}$

Mus musculus (2154): $\{0^{\{296\}}, 0.001047, 0.0014184, 0.001757, 0.0035151, 0.0041983, 0.0049098, 0.0069461, 0.0076774, 0.0083991, 0.010024, \dots, 0.88382, 0.88938, 0.8949, 0.90483, 0.91514, 0.92001, 0.92373, 0.92717, 0.94595, 0.96781, 1^{\{639\}}, 1.0101, 1.0188, 1.079, 1.0961, 1.1161, 1.119, 1.1471, 1.1528, 1.1544, 1.1554, \dots, 1.9834, 1.9854, 1.9874, 1.9892, 1.9908, 1.9909, 1.9931, 1.9936, 1.9963, 1.9983, 2^{\{276\}}\}$

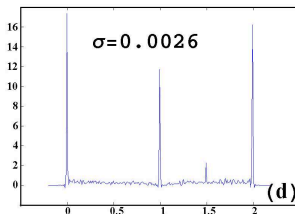
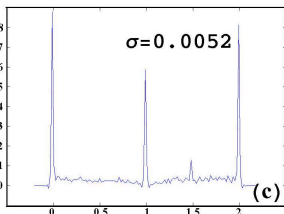
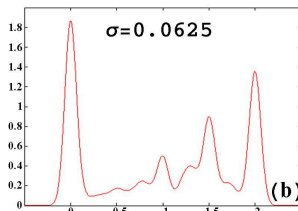
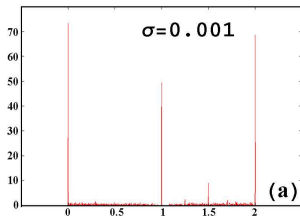
Spectrum plot

For a network G with n nodes, the eigenvalues set is $\{\lambda_i\}_{i=1}^n$.
We could use Gaussian kernel to smooth it:

$$\rho(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_G^2}} e^{-\frac{(x-\lambda_i)^2}{2\sigma_G^2}},$$

where σ_G is the smooth factor or bandwidth for this network.
Actually, smooth kernel (bandwidth) σ here is very important.

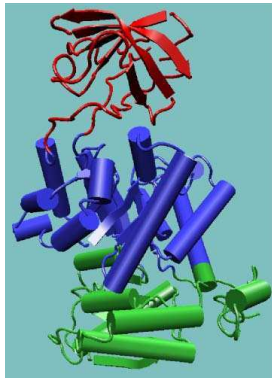
Spectrum plot



A reliable data-based bandwidth selection method for kernel density estimation, by S. J. Sheather and M. C. Jones

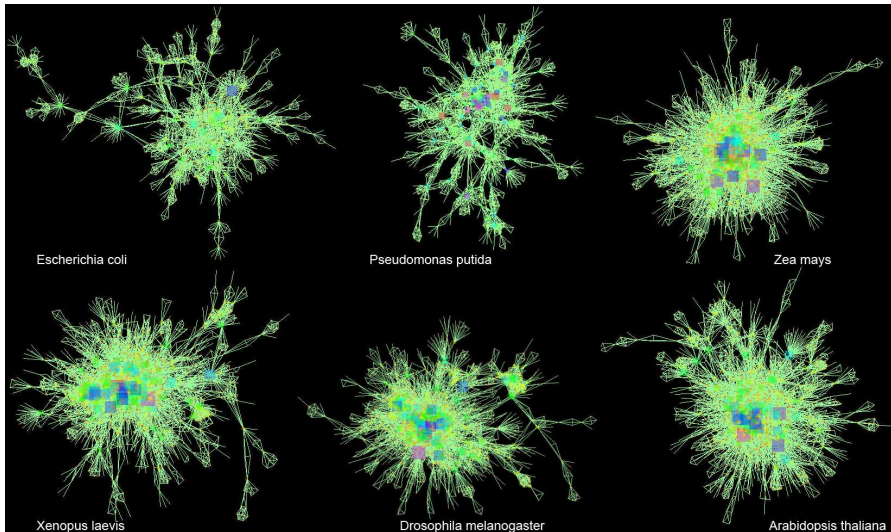
Kernel density estimation via diffusion, by Z. I. Botev, J. F. Grotowski and D.P. Kroese.

Protein and domain

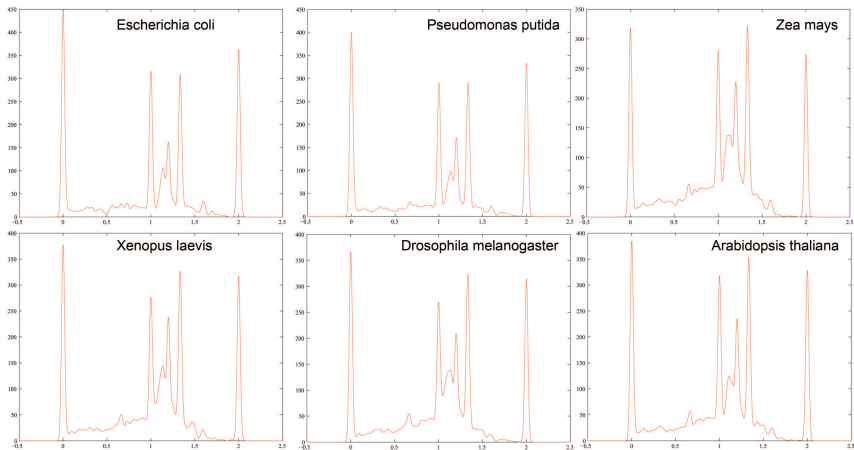


Proteins are organic compounds made of amino acids arranged in a linear chain and folded into a globular form. Domain is a part of protein sequence and structure that can evolve, function, and exist independently of the rest of the protein chain. Each domain forms a compact three-dimensional structure and often can be independently stable and folded. Many proteins consist of several structural domains.

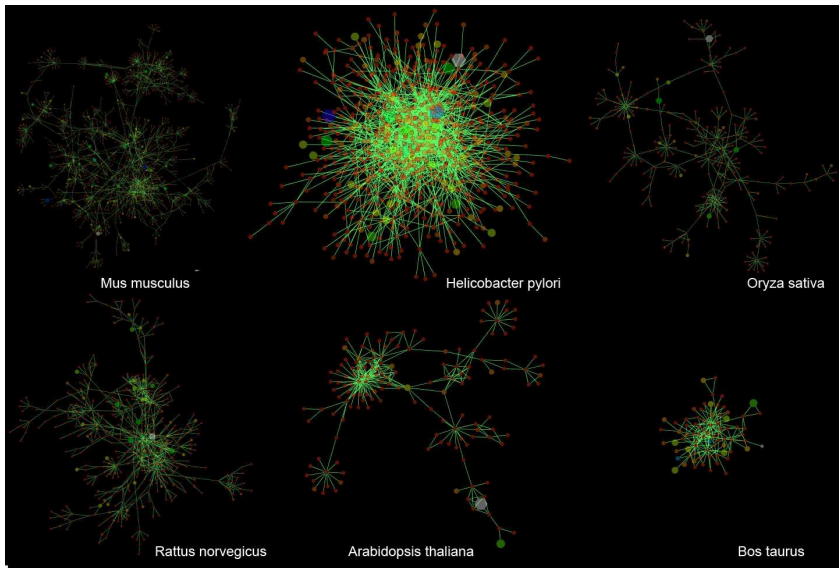
Domain co-occurrence networks



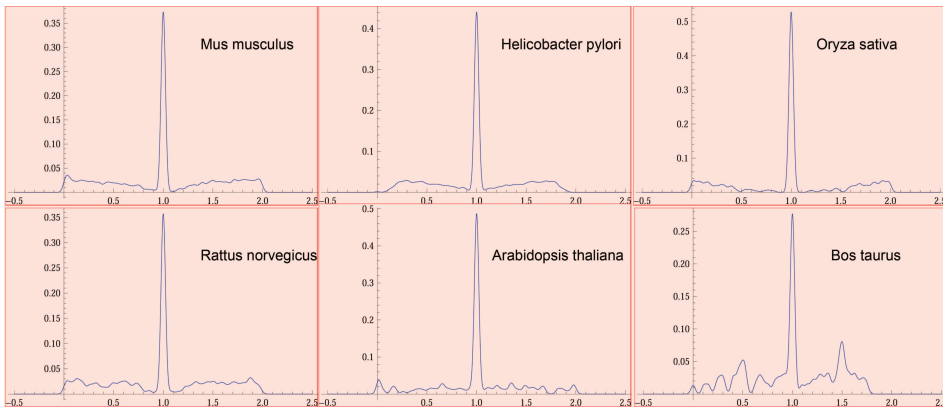
The spectrums of domain co-occurrence networks



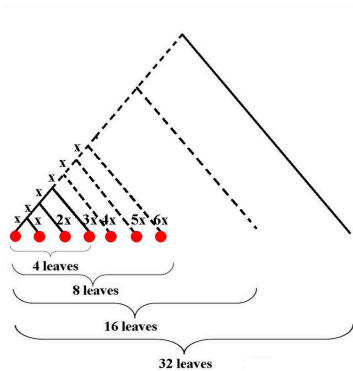
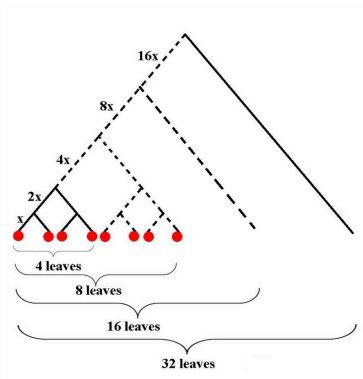
Protein interaction networks



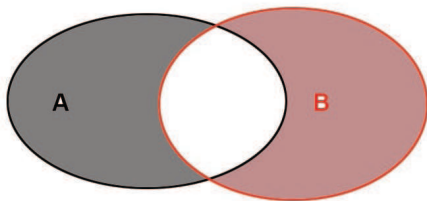
The spectrums of protein interaction networks



Reconstruct tree

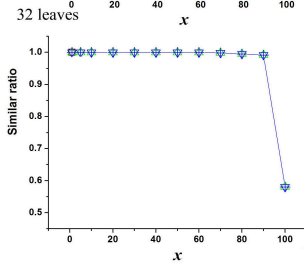
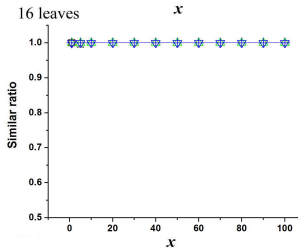
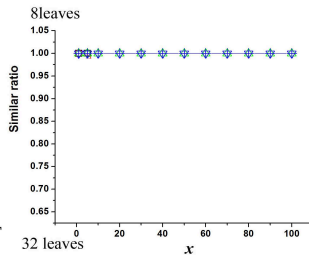
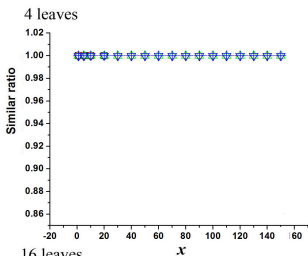


Edit distance



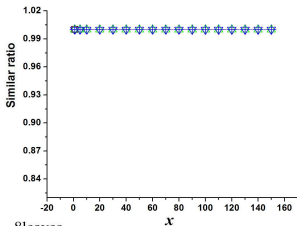
Edit distance d_e : The number (or ratio) of different edges between two graphs.

Edit distance

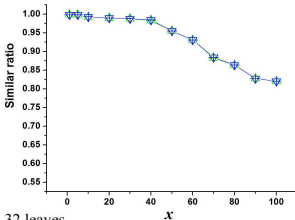


Edit distance

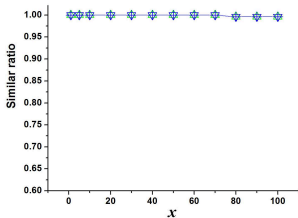
4 leaves



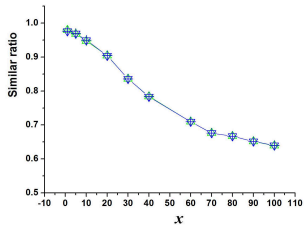
16 leaves



8leaves



32 leaves



Eigenvalues sets

Homo sapiens (5923): $\{0^{\{632\}}, 0.003413, 0.0058298, 0.006595, 0.010495, 0.015472, 0.015935, 0.016366, 0.019801, 0.020154, 0.020775, \dots, 0.95121, 0.95259, 0.95513, 0.96187, 0.96311, 0.96845, 0.9724, 0.97805, 0.98627, 0.99116, 1^{\{2235\}}, 1.0528, 1.0569, 1.0596, 1.0614, 1.0621, 1.0634, 1.0661, 1.0692, 1.0698, 1.0782, \dots, 1.9723, 1.9767, 1.9789, 1.9792, 1.9802, 1.984, 1.9842, 1.9843, 1.9932, 1.9935, 2^{\{567\}}\}$

Mus musculus (2154): $\{0^{\{296\}}, 0.001047, 0.0014184, 0.001757, 0.0035151, 0.0041983, 0.0049098, 0.0069461, 0.0076774, 0.0083991, 0.010024, \dots, 0.88382, 0.88938, 0.8949, 0.90483, 0.91514, 0.92001, 0.92373, 0.92717, 0.94595, 0.96781, 1^{\{639\}}, 1.0101, 1.0188, 1.079, 1.0961, 1.1161, 1.119, 1.1471, 1.1528, 1.1544, 1.1554, \dots, 1.9834, 1.9854, 1.9874, 1.9892, 1.9908, 1.9909, 1.9931, 1.9936, 1.9963, 1.9983, 2^{\{276\}}\}$

Bipartite matching distance

For two sequences $X = \{x_1, x_2, x_3, \dots, x_m\}$,
 $Y = \{y_1, y_2, y_3, \dots, y_n\}$, Bipartite matching distance is defined like this:

$$d_B(X, Y) = \sum_{i=1}^m \min_j |x_i - y_j| + \sum_{i=1}^n \min_j |y_i - x_j|. \quad (2)$$

Spectrum metric

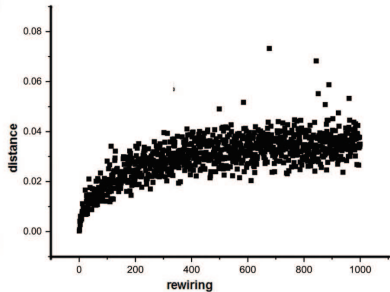
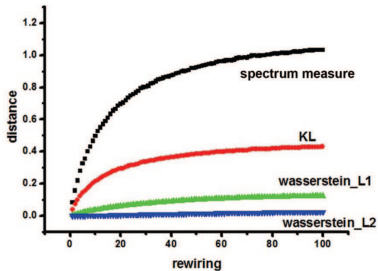
Spectrum metric for two networks G and G' is defined as follows: if normalized eigenvalues set of G is $\{\lambda_i\}_{i=1}^n$ and of G' is $\{\lambda_j\}_{j=1}^m$, then distance between them is :

$$d(G, G') = \int |\rho_i(x) - \rho_j(x)| dx,$$

$$\text{where } \rho_i(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_G^2}} e^{-\frac{(x-\lambda_i)^2}{2\sigma_G^2}}.$$

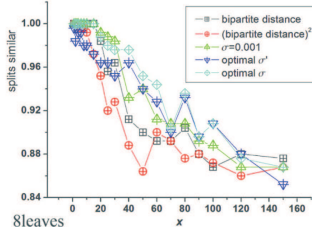
$$0 \leq d_{ij} \leq 2.$$

Distance with random graph

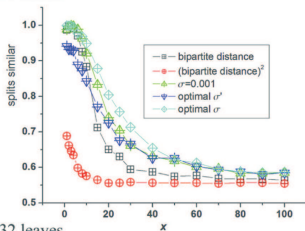


Experimental results

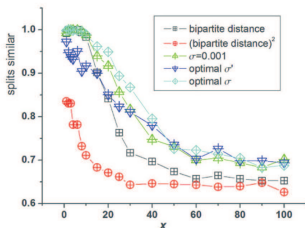
4 leaves



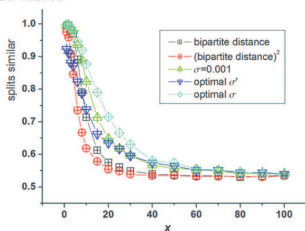
16 leaves



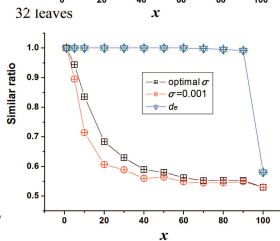
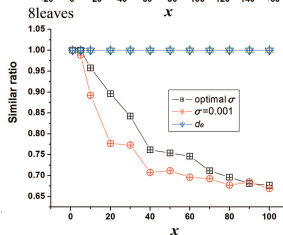
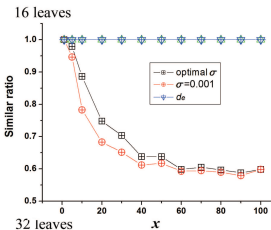
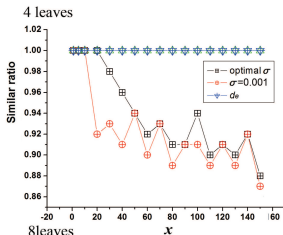
8 leaves



32 leaves

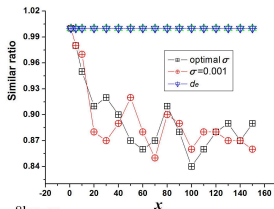


Experimental results

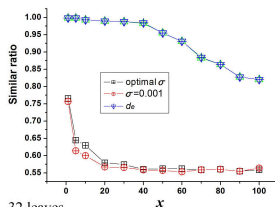


Experimental results

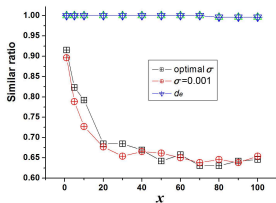
4 leaves



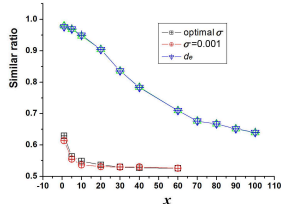
16 leaves



8leaves



32 leaves



Conclusion

- We proposed a method, aiming at compare two networks by their spectrums without identical information for each vertex. We try to get more structure information from spectrum, and hope to find more phylogenetic difference among networks by spectrum analysis.

Outlook

- We hope to improve reconstruction results using spectrums and apply this method on the biology networks.

Thanks to Prof. Peter Stadler, Prof. Juegen Jost.
Thanks to ...

Thank you very much for your attention!

