Who's this guy?

Knowledge Discovery in Databases through Complex Networks: application to phylodynamics

Luiz Max F. de Carvalho
Scientific Computing Programme (PROCC), Fiocruz
Pan American Center for Foot-and-Mouth Disease (PAHO/WHO)

WaFiS 2012

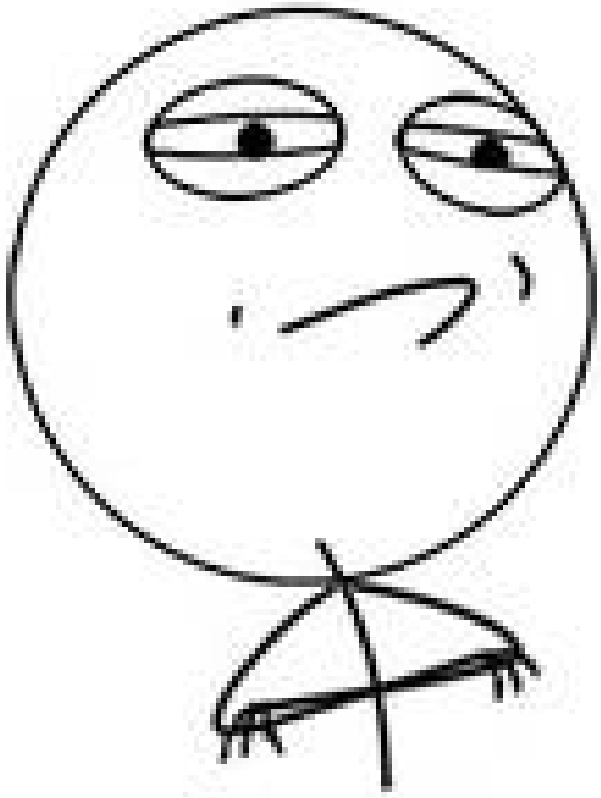
September 28, 2012

Outline

Knowledge Discovery in Databases (KDD)

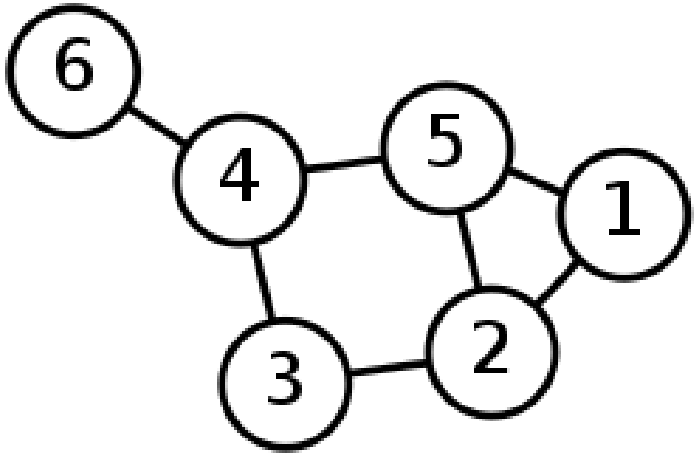
- Lots of data
- human brain ~~very~~ limited processing capacity
- Information $\rightarrow$ Knowledge
- Increasing number of molecular data (sequences, 3D structures, antigenicity,...)
- Is it possible to explore these databases to discover useful stuff?

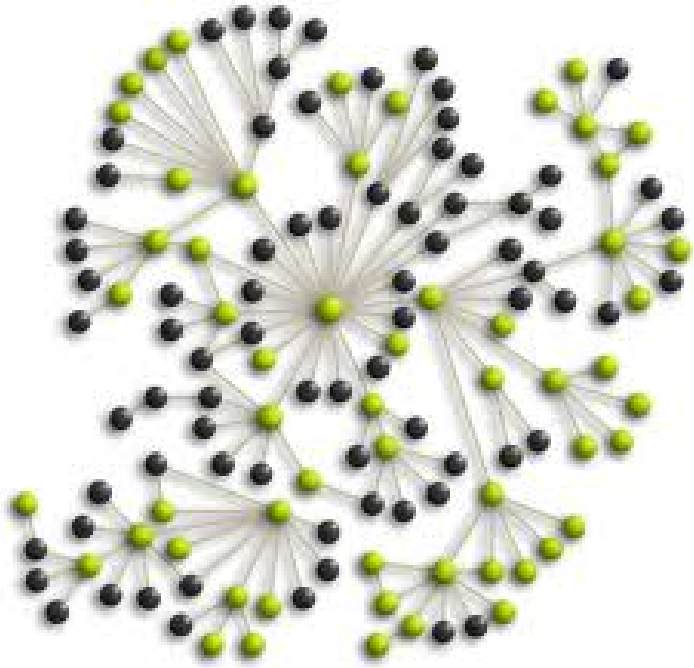
Well. . . Let's see



[We may use] Complex Networks

- Graphs $\rightarrow G = (V, E)$





Yeah, but how?

We can explore the "dynamic signature" of these Complex Networks, i.e., study and compare their structural properties. Some useful formulas:

- Clustering Coefficient $< c >$: $\frac{3 \times \#triangles}{\#triples}$
- Degree distribution $P_K = \sum_{K'=K}^{\infty} p_{K'}$
- Diameter: $max(d(i,j))$

Ok, Let's work then

1. Grab n sequences;
2. Create an $n \times n$ matrix using some kind of (normalized) distance (say, S);
3. For each $\sigma \in [0, 1]$ build $M(\sigma)$ such that:

$$m_{ij}(\sigma) = \begin{cases} 1 & \text{if } Sij > \sigma, \\ 0 & \text{if } S_{ij} < \sigma. \end{cases}$$

In a sense, we are transforming a single network in a family of networks.

Analysis

We shall explore the relationships between these networks: First, define a higher-order neighborhood indicator function, such that you binarize the adjacency matrix with regard the path length ℓ, obtaining a matrix $\hat{M} = \sum_{\ell=1}^{D} \ell M(\ell)$. Then

$$\delta(\alpha, \beta) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{\hat{m}_{ij}(\alpha)}{D(\alpha)} - \frac{\hat{m}_{ij}(\beta)}{D(\beta)} \right) \quad (1)$$

Evaluating $\delta(\sigma, \sigma + \Delta\sigma)$ can give some interesting insights.

Example 1: Chitin pathway phylogeny

Detecting Network Communities: An Application to Phylogenetic Analysis

Roberto F. S. Andrade[1], Ivan C. Rocha-Neto[2], Leonardo B. L. Santos[1,3], Charles N. de Santana[4], Marcelo V. C. Diniz[5], Thierry Petit Lobão[2], Aristóteles Goés-Neto[5], Suani T. R. Pinho[1], Charbel N. El-Hani[6]*

1 Institute of Physics, Federal University of Bahia, Campus Universitário de Ondina, Salvador, Bahia, Brazil, **2** Institute of Mathematics, Federal University of Bahia, Campus Universitário de Ondina, Salvador, Bahia, Brazil, **3** National Institute for Space Research, São José dos Campos, São Paulo, Brazil, **4** Mediterranean Institute of Advanced Studies, IMEDEA (CSIC-UIB), Esporles (Islas Baleares), Spain, **5** Department of Biological Sciences, State University of Feira de Santana, Feira de Santana, Bahia, Brazil, **6** Institute of Biology, Federal University of Bahia, Campus Universitário de Ondina, Salvador, Bahia, Brazil

- Proteins related to the chitin metabolic pathway from 1605 complete genomes;
- BLAST distances (which are **asymmetric**);
- Search for phylogenetic relationships

Example 1: Some results

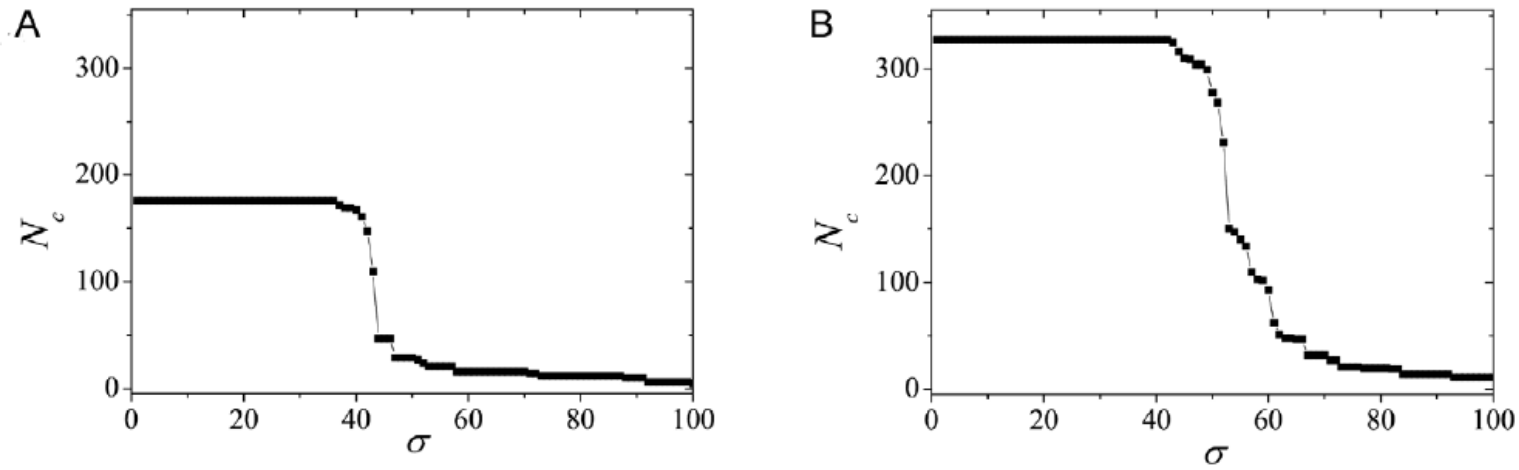


Figure 1. The size of the largest connected component (N_c) versus the threshold similarity σ: a) Acetyl; b) UDP.
doi:10.1371/journal.pcbi.1001131.g001

Example 1: Some more results

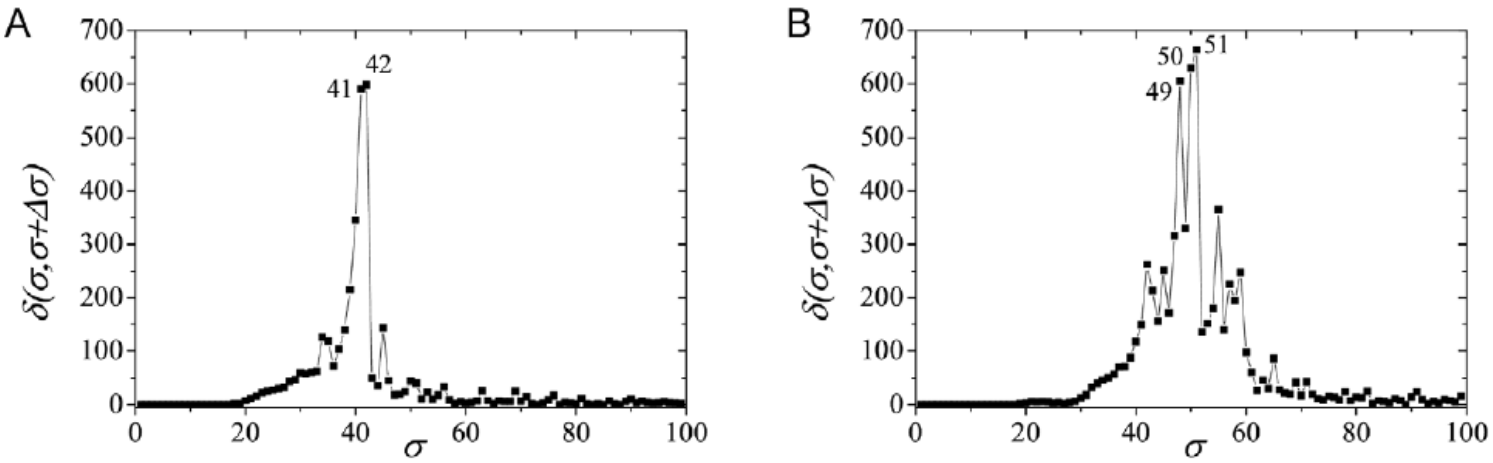
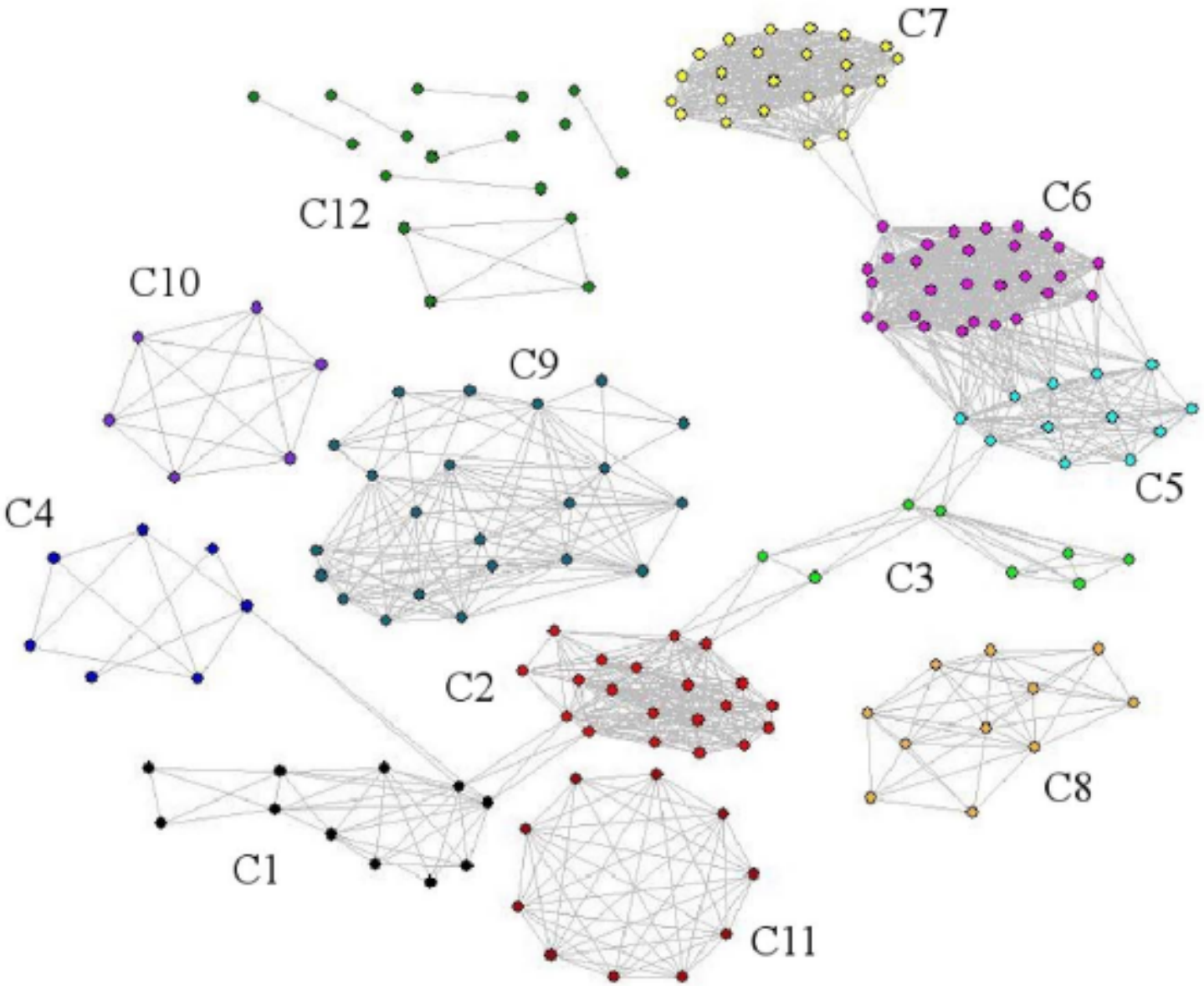


Figure 2. The distance $\delta(\sigma,\sigma+\Delta\sigma)$ between networks for successive similarities at the maximal value, with $\Delta\sigma=1$, in the case of: a) Acetyl at $\sigma=\sigma_{max}=42\%$; b) UDP at $\sigma=\sigma_{ma}=51\%$.
doi:10.1371/journal.pcbi.1001131.g002

Example 1: The expected Network(s)



Example 2: Foot-and-mouth disease virus in South America

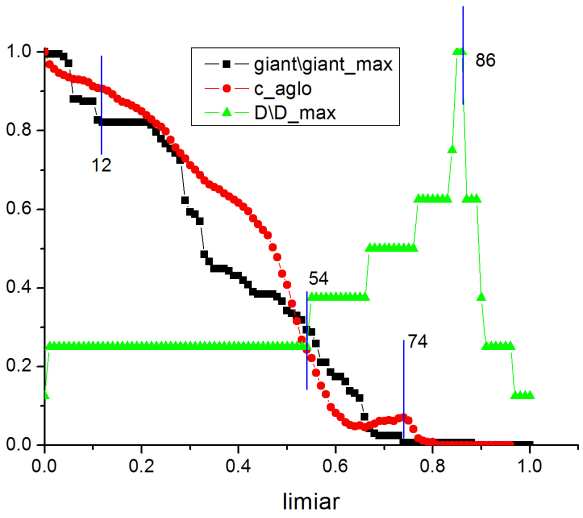
PHYLODYNAMICS OF FOOT-AND-MOUTH DISEASE VIRUS: A COMPLEX NETWORK APPROACH

Luiz Max Fagundes de Carvalho[1,3], *Leonardo Bacelar Lima Santos*[2], *Pedro Jeovah Pereira*[3], *Waldemir de Castro Silveira*[3]

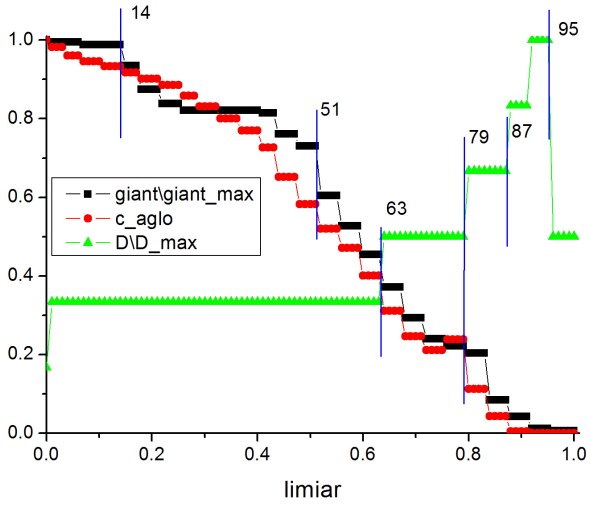
[1]Sector of Infectious Diseases Epidemiology – Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, luizepidemiologia@gmail.com
[2]National Institute os Space Research – INPE, São José dos Campos – SP , Brazil, santoslbl@gmail.com
[3]Laboratory of Bioresources – Pan American Foot-and-mouth Disease Center (PANAFTOSA) – Pan American Health Organization (PAHO), Duque de Caxias – RJ, Brazil, silveiraw@paho.org, ppereira@paho.org

- S was built with phylogenetic (TN93) distances for NT and JTT distances for AA;
- Try to make sense of a somewhat big data set (167 seqs);
- Extract some nice patterns;

Indexes $\times$ σ

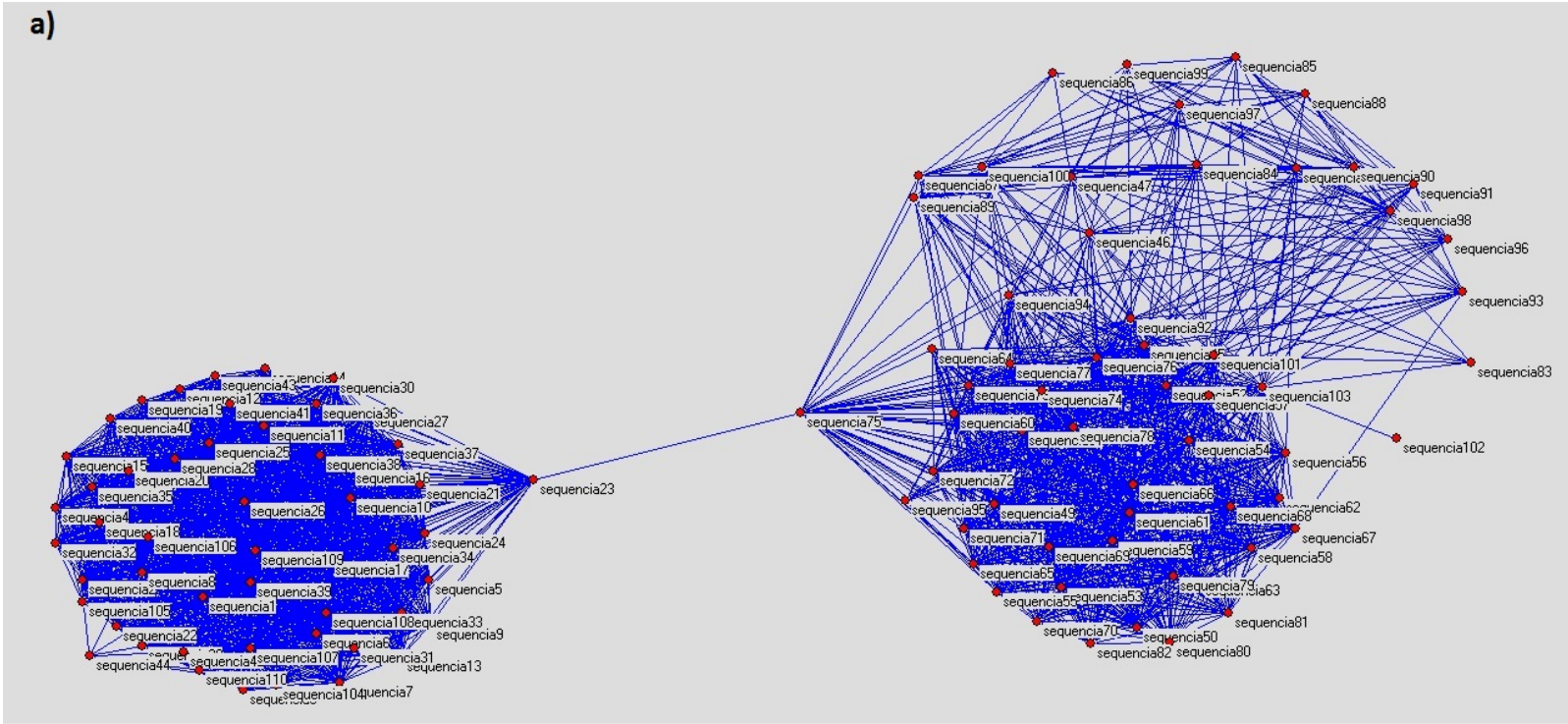


(a)

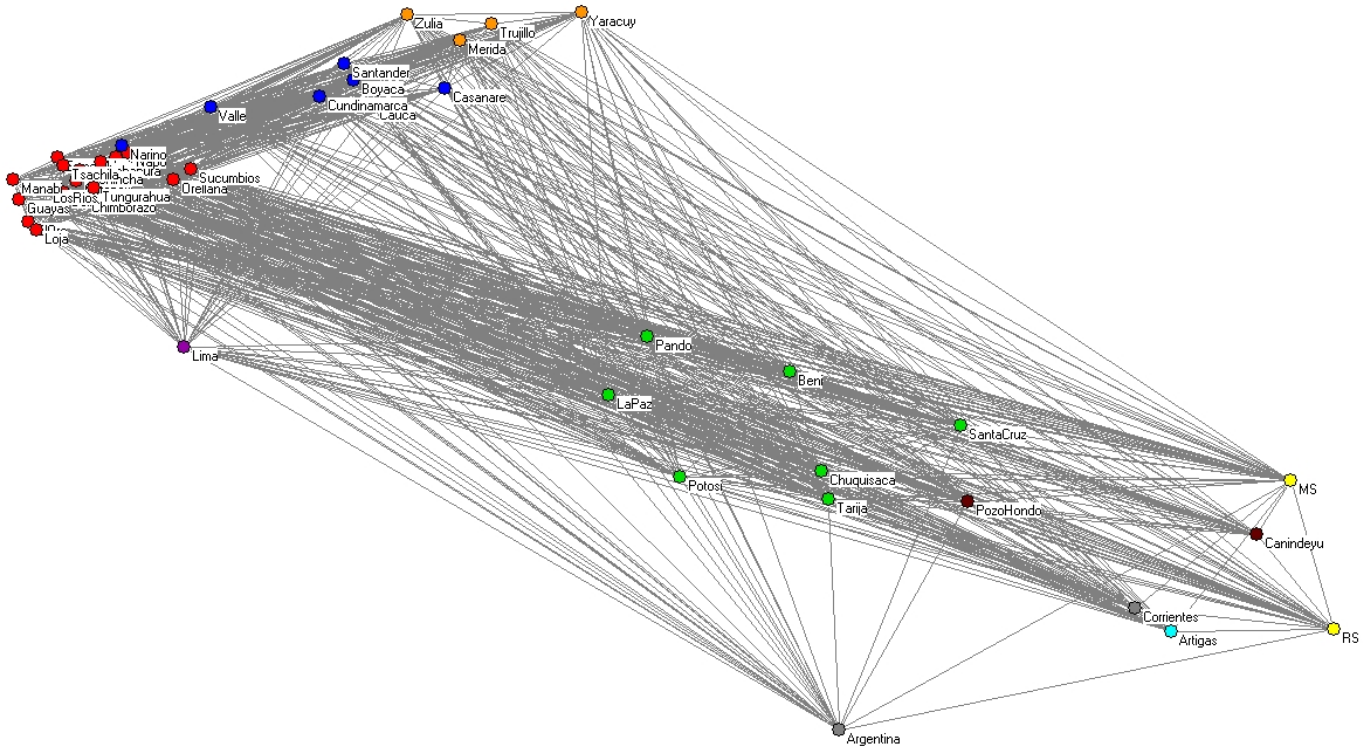


(b)

A nice network



Some more developments



Related Work

- Identify transmission clusters (HIV, HCV) (Lewis et al, 2008,Plos Medicine)
- Explore scale-free behavior in phylodynamics (Shiino, 2012, Frontiers in Microbiology)

Future Directions

- Explore the spatial aspect in the construction of S Maybe $\hat{S} = \mu + S(G)^{\alpha}$
- Power law analysis
- Implement assortativity
- Suggestions...

Thank You!