Visualization and Data Integration Techniques in Systems Biology.

Georgios Pavlopoulos [1*], Charalampos Moschopoulos[2], Evangelos Pafilis[3] and Reinhard Schneider[1]

[1] European Molecular Biology Laboratory, Heidelberg, Germany [2] Biomedical Research Foundation Academy of Athens, Greece [3] University of Crete, Heraklion.
[*]Correspondence to: pavlopou@embl.de

Motivation: Bioinformatics has evolved and expanded continuously over the past four decades and has grown into a very important bridging discipline in life science research. The quantities of data obtained by new high-throughput technologies are vast and biological data repositories are growing exponentially in size. This study aims to demonstrate different approaches of collecting, analyzing and integrating heterogeneous data. The interest was mainly focused on implementing visualization tools that are able to visualize networks in both 2D and 3D dimensions. The visualization techniques were applied to identify protein complexes and visualize interactions between human GPCRs, G-proteins and effectors [1, 2]. Text-mining machineries were also developed to analyze data and provide informative networks about bioentities and their interactions. The projects mainly developed and described in this abstract are the Arena3D [3], Medusa [4], OnTheFly [5], GIBA [6, 7] and jClust [8].

Arena3D [3] is a tool that introduces a new concept of staggered layers in 3D space and is suitable for multi-layered graph visualization. Related data - such as proteins, chemicals, or pathways can be grouped onto separate layers and arranged via layout algorithms to make the network more informative. Arena3D is rich in clustering algorithms and it is suitable to visualize large scale networks consisting of heterogeneous data.

Medusa [4] is a tool for visualization and clustering analysis of biological networks in 2D. It is highly interactive and it supports weighted and multi-edged directed and undirected graphs where each edge between two bioentities can represent a different biological concept. Medusa is currently enriched with a variety of layout and clustering methods to bridge the gap between analysis and visualization

GIBA [6] and ***jClust*** [8] are two applications which utilize an important collection of clustering and clique finding algorithms. These algorithms are k-Means [9], Affinity Propagation [10], Bron-Kerbosch [11], Restricted Neighborhood Search Cluster Algorithm [12], Markov Clustering [13] and Spectral Clustering [14]. Filters that can be applied on the clustering results are utilized by 1) haircut, 2) outside-inside, 3) best neighbors and 4) density control operations. GIBA was used to predict protein complexes from PPI networks. GIBA implements a new two step methodology: In the first step, the protein interaction network is clustered by the Markov Clustering algorithm [13] or the Restricted Neighborhood Search Clustering Algorithm [RNSC] [12]. In the second step, the clustering results are filtered to derive the final candidate protein complexes. Clusters can be visualized by a new version of Medusa.

OnTheFly [5] is a web-based application that extends the Reflect [15] functionality to identify gene, protein and chemical names in Microsoft Word, Excel, Power Point, PDF and plain text format files and provide informative popup windows for each of the bioentities. It supports the extraction of protein-protein interactions generated by the STITCH [16] database and it is able to generate information tables showing additional information about the tagged bioentities like IDs, names and descriptions.

Acknowledgements: The authors wish to thank in advance all HSCBB09 participants for inspiration and future collaborations.

References

1. Theodoropoulou MC, Bagos PG, Spyropoulos IC, Hamodrakas SJ: **gpDB: a database of GPCRs, G-proteins, effectors and their interactions**. *Bioinformatics (Oxford, England)* 2008, **24**(12):1471-1472.
2. Elefsinioti AL, Bagos PG, Spyropoulos IC, Hamodrakas SJ: **A database for G proteins and their interaction with GPCRs**. *BMC bioinformatics* 2004, **5**:208.
3. Pavlopoulos GA, O'Donoghue SI, Satagopam VP, Soldatos TG, Pafilis E, Schneider R: **Arena3D: visualization of biological networks in 3D**. *BMC Syst Biol* 2008, **2**(1):104.
4. Hooper SD, Bork P: **Medusa: a simple tool for interaction graph analysis**. *Bioinformatics* 2005, **21**(24):4432-4433.
5. Pavlopoulos GA, Pafilis E, Schneider R, Kuhn M, Sean DH: **OnTheFly: A Tool for automated document-based text annotation, data linking and network generation.** *Bioinformatics* 2009, **10**.
6. Charalampos N. Moschopoulos, Georgios AP, Reinhard S, Spiridon DL, Sophia K: **GIBA: A clustering tool for detecting protein complexes**. *BMC Bioinformatics* 2009.
7. Charalampos. N. Moschopoulos, Georgios. A. Pavlopoulos, Spiridon. D. Likothanassis, Kossida S: **An enhanced Markov clustering method for detecting protein complexes**. *8st IEEE International Conference on Bioinformatics and Bioengineering* 2008.
8. Pavlopoulos GA, Moschopoulos CN, Hooper SD, Schneider R, Kossida S: **jClust: a clustering and visualization toolbox**. *Bioinformatics* 2009, **25**(15):1994-1996.
9. MacQueen JB: **Kmeans Some Methods for classification and Analysis of Multivariate Observations**. In: *5-th Berkeley Symposium on Mathematical Statistics and Probability: 1967; Berkeley*: University of California Press; 1967: 281-297.
10. Frey BJ, Dueck D: **Clustering by passing messages between data points**. *Science* 2007, **315**(5814):972-976.
11. Kerbosch CB-J: **Algorithm 457: finding all cliques of an undirected graph**. *ACM Press* 1973, **16**(9):575 - 577.
12. King AD, Przulj N, Jurisica I: **Protein complex prediction via cost-based clustering**. *Bioinformatics* 2004, **20**(17):3013-3020.
13. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic acids research* 2002, **30**(7):1575-1584.
14. Paccanaro A, Casbon JA, Saqi MA: **Spectral clustering of protein sequences**. *Nucleic acids research* 2006, **34**(5):1571-1580.
15. Pafilis E, O'Donoghue. SI, Jensen. LJ, Kuhn. M, Horn. H, Brown. NP, Schneider. R: **Reflect: Augmented Browsing for the Life Scientist. http://reflect.ws**. *Nature Biotechnology* 2009.
16. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P: **STITCH: interaction networks of chemicals and proteins**. *Nucleic acids research* 2008, **36**(Database issue):D684-688.