Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Web Server**

NASCENT: An automatic protein interaction network generation tool for non-model organisms

Daniel Banky [1, 2], Rafael Ordog [1, 2], Vince Grolmusz [1, 2, *]

[1]Protein Information Technology Group, Eotvos University, H-1117 Budapest, Hungary, [2]Uratim Ltd. H-4400 Nyiregyhaza, Hungary;
Vince Grolmusz - E-mail: grolmusz@cs.elte.hu; *Corresponding author

received March 19, 2009; revised April 02, 2009; accepted April 07, 2009; published April 24, 2009

Abstract:
Large quantity of reliable protein interaction data are available for model organisms in public depositories (e.g., MINT, DIP, HPRD, INTERACT). Most data correspond to experiments with the proteins of *Saccharomyces cerevisiae, Drosophila melanogaster, Homo sapiens, Caenorhabditis elegans, Escherichia coli* and *Mus musculus*. For other important organisms the data availability is poor or non-existent. Here we present NASCENT, a completely automatic web-based tool and also a downloadable Java program, capable of modeling and generating protein interaction networks even for non-model organisms. The tool performs protein interaction network modeling through gene-name mapping, and outputs the resulting network in graphical form and also in computer-readable graph-forms, directly applicable by popular network modeling software.

Availability: http://nascent.pitgroup.org

Keywords: network; tool; protein; interaction; model

Background:
Experimentally validated, high quality protein-protein interaction data are deposited in numerous information sources on the Web, for example in databases like MINT **[1],** HPRD **[2],** DIP **[3]** and IntAct **[4]**. Most of the data were acquired for popular model organisms, like Saccharomyces cerevisiae **[5],** Drosophila melanogaster, Caenorhabditis elegans **[6],** Escherichia coli, Mus musculus and Homo sapiens **[7, 8, 9].** Numerous other organisms of importance are completely missing from these depositories, or only very little data are deposited publicly (e.g., Mycobacterium tuberculosis).

Modeling protein interaction networks for organisms poorly represented in the large depositories is an important task. In the present work we describe a web-based tool, called NASCENT, capable of automatically modeling of protein interaction network data from the rich experimental data source deposited in IntAct **[4]**. In NASCENT, the user is allowed to designate a source organism and the target organism. The source organism, preferably one of those with lots of deposited protein-protein interaction data, is used for input in modeling the interactions in the target organism. The organisms need to be identified by NCBI taxonomy ID's; an autocompleting tool helps the user to find this ID from the scientific Latin name of the species.

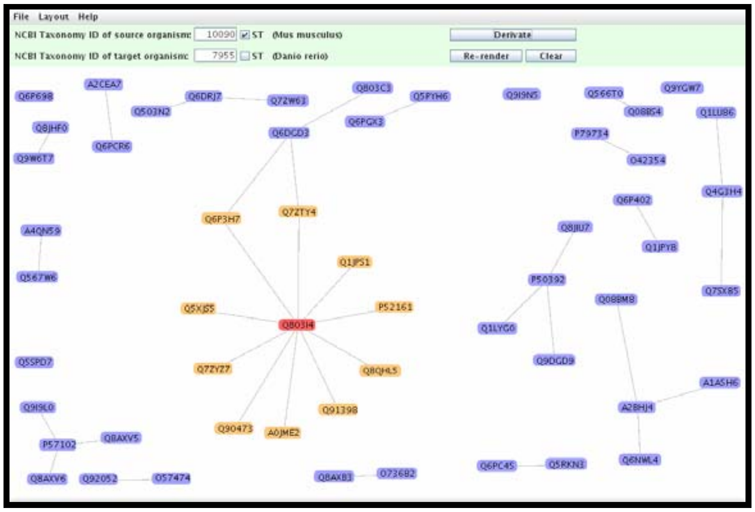
The mapping of the interactions is done by corresponding genes of the expressed proteins of the two organisms, as described in the Methodology section. NASCENT applies the Swiss-Prot database **[10]** for performing the mapping of the gene names of different organisms. The NASCENT is scalable tool for integrating the constantly updated source database (IntAct **[4]** and the mapping database (UniProtKB/Swiss-Prot **[10]:** it applies weekly updates to the internal database, queried by the tool.

The graphical interface of Nascent is intended to yield a quick overview of the network generated. The nodes are labeled by the UniProt primary accession numbers of the proteins **[10],** and the drawing method of the graph can be selected as one of seven layouts (i.e., random, force-directed, Fruchterman-Reingold force directed, node-link tree, balloon tree, radial tree and circle layouts). For the easier navigation through complex networks, when a mouse cursor moves over a protein-code, then its color will be changed to red, while their neighbors will be yellow. Network nodes can also be moved around and can be grouped easily with the mouse.

The NASCENT tool is capable of creating networks in SIF, text and GraphML output for using in other network tools. JPEG export can also be chosen at http://nascent.pitgroup.org. A faster, downloadable self-standing Java program is also available there.

Methodology:
The network construction algorithm is summarized on Figure 2. When the user designates a species, the network of that species is retrieved from the local, regularly updated mirror of the IntAct database **[4].** If the user checks the box requiring the inclusion of the phylogenic sub-tree on the source side, then the protein interaction network data of all the descendent subspecies of the source organism will also be included in the list. That forms the local copy of the interaction network of the source species.

Next, the protein-gene correspondence is computed, using the UniProt database **[10].** The target organism's genes and proteins are chosen from the UniProt database **[10]** as the next step. If the user checks the box requiring the inclusion of the phylogenic subtree on the target organism, then the dictionary of the genes and protein accession codes of the descendent subspecies of the target organism will also be included in the list. Next the proteins of the source and the target organisms are corresponded according to the gene names; we call this step gene mapping on Figure 2.



NCBI Taxonomy ID of source organism: 10090 ☑ ST (Mus musculus)

NCBI Taxonomy ID of target organism: 7955 ☐ ST (Danio rerio)

Figure 1: Screenshot of the NASCENT tool: The network of *Danio rerio* generated by gene-correspondence from *Mus musculus*. Vertices are labeled by primary accession numbers from UniProt. In this screen-shot, the red node is chosen and its neighbours are colored yellow.

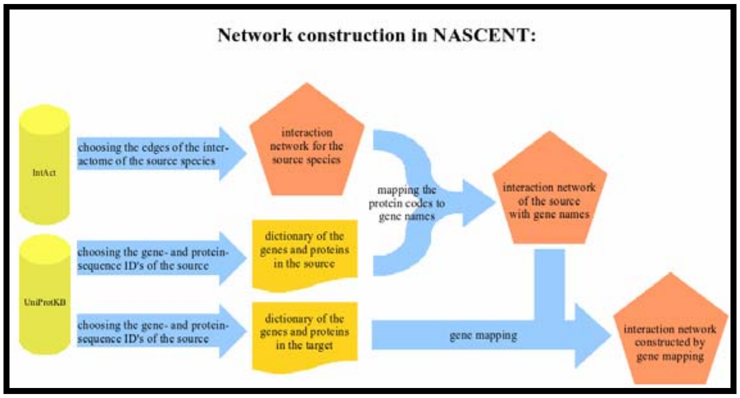

Network construction in NASCENT:
IntAct
UniProtKB
choosing the edges of the inter-actome of the source species
interaction network for the source species
choosing the gene- and protein-sequence ID's of the source
dictionary of the genes and proteins in the source
choosing the gene- and protein-sequence ID's of the source
dictionary of the genes and proteins in the target
mapping the protein codes to gene names
interaction network of the source with gene names
gene mapping
interaction network constructed by gene mapping

Figure 2: Flow-chart of the network construction by NASCENT.

Multiple correspondences are handled as follows: Suppose that the source organism is X and the target organism is Y. If in the source organism X the interaction edge A-B is present, connecting proteins A and B, and protein A corresponds to proteins A1, A2 and A3 in organism Y, and protein B in X corresponds to protein B1 in Y, then all the three edges A1-B1, A2-B1 and A3-B1 will be added to the network of the target. The graphical user interface was made by using the Prefuse toolkit http://prefuse.org. The Java installer application chosen is the install4j of ej-technologies http://www.ej-technologies.com.

Caveats:

NASCENT will return an empty graph for target organisms with very few data in the SwissProt database. For example, if one try to generate the network of *Canis lupus* from the network of the *Mus musculus*, and the "Include subtree" box is unchecked next to the target organism, then the result graph will be empty, since there is relatively little data on the grey wolf in Uniprot. However, if we check the box "Include sub-tree" box, then all the data of the subspecies will be screened, including those of the domestic dog, so the graph will not be empty. It is recommended to check that box if NASCENT returns an empty graph. Presently, only NCBI Taxonomy ID's of species can be entered, the codes of subspecies will not generate outputs; except in the case when all the subspecies of a given species is screened as in the example above. Note, that generating large force-directed layouts is resource-hungry.

Utility:

The most useful application is generating protein interaction networks for important, but non-model organisms, with lots of data in Uniprot. The network generated can be exported into popular graph drawing and network analyzing software, since we offer text, SIF and GrapML formatted output. For fast review the Java applet should be chosen by clicking "Show Graph" on the Results Page. There the nodes can be moved around, and the neighbors are marked automatically, by the touch of the mouse.

Acknowledgement:
This research was partially supported by the NKTH grants OMFB-01295/2006 and OM-00219/2007 and OTKA grant NK-67867.

References:

[1] A. Chatraryamontri, *et al. Nucleic Acids Res* **35**(Database issue), D572–D574 Jan (2007).

[2] T. S. K. Prasad, *et al. Nucleic Acids Res* **37**(Database issue), D767–D772 Jan (2009).

[3] L. Salwinski, *et al. Nucleic Acids Res* **32**(Database issue), D449–D451 Jan (2004).

[4] S. Kerrien, *et al. Nucleic Acids Res* **35**(Database issue), D561–D565 Jan (2007).

[5] A. X. C. N. Valente, *et al. Proc Natl Acad Sci U S A* **106**(5), 1490–1495 Feb (2009).

[6] S. Li, *et al. Science* **303**(5657), 540–543 Jan (2004).

[7] J. F. Rual, *et al. Nature* **437**(7062), 1173–1178 Oct (2005).

[8] M. E. Futschik, *et al. Bioinformatics* **23**(5), 605–611 Mar (2007).

[9] G. Chaurasia, *et al. Nucleic Acids Res* **37**(Database issue), D657–D660 Jan (2009).

[10] UNIPROT Consortium, *Nucleic Acids Res* **37**(Database issue), D169–D174 Jan (2009).

Edited by P. Kangueane
Citation: Grolmusz *et al***,** Bioinformation 3(8): 361-363 (2009)