Published online 31 July 2009

Nucleic Acids Research, 2009, Vol. 37, No. 12 ***1–5***
doi:10.1093/nar/gkn000

NaviCell Web Service for Network-based Data Visualization

Eric Bonnet [1,2,3], Eric Viara [4], Inna Kuperstein [1,2,3], Laurence Calzone [1,2,3], David PA Cohen [1,2,3], Emmanuel Barillot [1,2,3], Andrei Zinovyev [1,2,3*]

[1]Institut Curie, 26 rue d'Ulm, 75248 Paris, France, [2]INSERM U900, 75248 Paris, France, [3]Mines ParisTech, 77300 Fontainebleau, France, [4]Sysra, 91330 Yerres, France.

Received January 1, 2009; Revised February 1, 2009; Accepted March 1, 2009

ABSTRACT

We developed NaviCell Web Service for network-based visualization of "omics" data, using Google Maps API for browsing large maps of molecular interactions and RESTfull web service with available Python binding to automate the data visualization tasks. NaviCell Web Service implements a number of data visual representation methods including the novel map staining technique which allows grasping large-scale trends in numerical values (such as whole transcriptome) projected on top of the pathway map. We provide several case studies using the pathway maps created by different research groups in which data visualization serves for getting new insights into molecular mechanisms involved in systemic disease progression such as cancer and neurodegenerative diseases.

INTRODUCTION

Biology is a scientific discipline deeply grounded in visual representations serving for communicating results and ideas. Nowadays, there is a strong incentive to provide *interactive* web-based visual representations, where users can easily play with different options to tweak and adapt them. Modern molecular biology is particularly demanding for development of new tools representing in a meaningfull way numerous "omics" data collected in large-scale experiments. Visualizing quantitative "omics" data in the context of biological networks provides insights into the molecular mechanisms in healthy tissues and disease (1, 2). This is one of the most demanded features of existing pathway databases such as KEGG PATHWAY and Reactome (3, 4). To answer this need, many tools have been developed allowing mapping "omics" data on top of the biological networks (5, 6, 7, 8). These tools use the content of existing pathway databases to retrieve the pathway information through dedicated APIs. In addition, some pathway databases provide graphical user interfaces (GUI) to perform the task of data visualization, on top of their interactive pathway maps.

When a user is faced with necessity to visualize the "omics" data on top of biological networks, there are currently two options available: either to use GUI of a suitable pathway database, which can be a tedious manual work, or use one of the available standalone or web-based tools for creating static images of colored pathways which are not interactive anymore and do not allow browsing the molecular interactions. Currently, there exist no well-developed APIs allowing programmatically apply data visualization on top of biological networks such that the omics data can be browsed simultaneously with the pathway information.

Available network-based methods for data visualization have several common limitations. First, most of them do not provide possibility of data abstraction, allowing to visualize coarse-grained trends in the omics data. This is needed if, for example, there is a wish to visualize the whole transcriptome of a cell or a group of samples on top of a large network, representing a big part of cellular interactome. Current methods usually attach some elements of standard scientific graphics (gradient color, heatmaps, barplots) to the individual elements of pathway maps, which makes the visualization hardly readable at higher levels of zoom as well as the map content itself. Second, most of data visualization tools are specific to a particular pathway database structure, format and the way of graphical pathway representation. This limits the use of data visualization on top of user-defined pathway maps. Third, APIs for programmatic web-based SBGNdata visualization are practically non-existent.

To fill this niche, we've developed NaviCell Web Service tool. It allows visualizing omics data via GUI and flexible API using standard and advanced methods of data visualization including a possibility to perceive the data structure at different network scales. The NaviCell Web Service allows exploiting pathway maps created by users themselves or from existing databases, including very large maps (for example, representing the content of a whole pathway database). The NaviCell Web Service for Network-based Data Visualization is a combination of 1) user-friendly NaviCell JavaScript-based web interface (9), allowing browsing very large maps of biological networks using Google Maps API using *semantic zooming* principle and visualize "omics" data on top of them (http://navicell.curie.fr; 2) large collection of maps available for data visualization including the Atlas of Cancer Signaling Network (ACSN) (http://acsn.curie.fr and other detailed network maps created by many research groups for a particular subject (Toll-like receptor signalling, EGFR and mTOR pathways, mast cell activation, dendritic cells, iron metabolism, Alzheimer disease and others); 3) REST API allowing programmatic use of all data visualization functions and manipulating the web interface, with available Python binding and upcoming R and Java bindings.

*To whom correspondence should be addressed. Tel: +33 (0)1 56 24 69 89; Fax: +33 (0)1 56 24 69 11; Email: andrei.zinovyev@curie.fr

Quick live example which *automatically* uploads a sample dataset from TCGA database and illustrates all basic data visualization capabilities of NaviCell Web Service is provided at http://navicell.curie.fr/pages/nav_web_service.html Sample Python code for quick start on using NaviCell Web Service API is available from the same page in "Python API and data files" section.

ALGORITHMS AND SOFTWARE

Implementation

We conceived the NaviCell Web Service for Network-based Data Visualization as a significant development of NaviCell tool (9) allowing users to visualize and analyze different types of "omics" data. The web service is implemented at two interconnected functional levels. The first level corresponds to the interactive dialog-based use of the web interface to load and visualize data. This level is implemented as a specific set of JavaScript functions linked to the menu located in the right-hand panel of the NaviCell web interface. The second level corresponds to a programmatic usage of the service, which allows users to write the code that will communicate with the NaviCell *server* in order to automate all the visualization operations. This functional level is implemented as a RESTful web service, using standard web protocol HTTP operations and data encoding (JSON) to perform all the necessary operations (10). RESTful web services are very popular and have the advantage of being lightweight, efficient and simple to use. They also have the benefit of relatively ease of implementation in many different programming languages and software packages. At the moment we have a full API ready for Python, and APIs for R and Java in development. These APIs will facilitate the integration of NaviCell Web Service in other applications: among them, there exists an ongoing work on providing the data visualization services through the Garuda Alliance web platform, which aims to be a one-stop service for bioinformatics and systems biology (11) and GeneSpring software (12). Figure **??** is summarizing the NaviCell *server* architecture and information flow.

Input data types

The pathway maps which can be used for NaviCell Web Serivce data visualization are to be prepared using CellDesigner tool(13), which is widely used in systems biology community and implements the Systems Biology Graphical Notation standard (14) for representing biological networks. These pathway maps are converted into interactive Google Maps-based web-interfaces via NaviCell tool(9).

NaviCell data visualization web service is able to process several types of "omics" data. The complete list of data types that are currently accepted for input is provided in Table 1. The different biological data types are mapped into various internal representations that determine what methods of data visualization can be applied.

Table 1. List of input biological data types for NaviCell Web Service data visualization. The first column lists the types as they appear in the NaviCell Web Service interface. The second column lists the internal data representations that are used to determine what type of data visualization can be applied for a selected data type.

"Omics" data type	Internal representation
mRNA expression data	continuous
microRNA expression data	continuous
Protein expression data	continuous
Discrete copy number data	discrete ordered
Continuous copy number data	continuous
Mutation data	discrete unordered
Gene list	set

For instance, a mRNA expression data matrix is associated to a "continuous" numerical internal representation. Thus, if the user is choosing to display this data with a heatmap, mapping to a color gradient will applied by default, with a possibility for modifying the default setting. On the other hand, when a matrix with discrete copy-number data is loaded, it is associated with a "discrete unordered" **AZ: WHY UNORDERED, should be ORDERED???** internal representation for which a specific color palette is applied for visualization, with a distinct color associated to each discrete copy-number state.

The input format for data sets is standard tab-delimited text files, with rows representing genes or their products and columns representing samples (or experiments, or time points). Genes in the first column should be labeled by their standard HUGO (HGNC) gene symbols, that will be associated to the different entities (genes, proteins, complexes) on the pathway map.

Users can also upload sample annotation files to specify how samples can organized into meaningfull groups (e.g., disease vs control), as a simple tab-delimited text file with sample names in rows and annotation fields as columns. Then, an appropriate method will be used to summarize the values of all the samples contained in a group, defined by the internal data representation. For instance, expression values for a group of samples is averaged by default for the "continuous" data type, but the user will have a possibility to change the default method and use the median or the maximum value in the group, etc. For mutation data ("discrete unordered" internal data type), taking the average does not make sense and instead the grouping method by default is "at least one element of the group is mutated".

Graphical data representations

We have included in the NaviCell Web Service for network-based data visualization several methods for graphical representation of molecular data. Some of them are standard and broadly used in molecular biology (heatmaps and barplots), while others (glyphs and map stainings) have not been previously employed (to our knowledge) in the context of network-based data visualization.

- **Simple markers** are pictograms (similar to the ones used in Goggle Maps to indicate the geographical locations) are drawn on the pathway map at the location

of a given molecular species (protein, gene, complex, phenotype). They are used to display the results of a search performed by the user or to display a list of genes of interest uploaded on the map.

- **Heatmaps** display individual values of a data matrix as colors. They are often used in molecular biology, in particular, for displaying expression data. In NaviCell Web Service, heatmaps can be used to display continuous data types such as expression values, or discrete data such as copy-number or mutation values. The user can arrange the display to have several samples or group of samples displayed in several data sets (for example, for showing simultaneously, gene promoter methylation and expression values).

- **Barplots** are charts with rectangular bars proportional to the values they represent. On NaviCell Web Sevice interface, they are colored according to the data value. Barplots can be efficient to visually distinguish numerical values between two or more groups of conditions (e.g. disease vs control).

- **Glyphs** are graphical representations using basic geometrical shapes (triangle, square, rectangle, diamond, hexagon). The user can specify which datasets to use to map on the shape, the size and the color of the glyph. This type of representation is particularly useful to visually combine different types of data. For example, one might consider using matched data for gene copy number and expression. In this case, the shape of the glyphs could be assigned to the copy-number values, while the expression data would be used for the color of the glyph. Like this, users can quickly appreciate the two data types values at the same time on the glyph.

- **Map staining** is a novel network-based data visualization method where background areas (or territories) around each molecular entity are colored according to the data value associated to this entity. The area occupied by a particular map element is defined as the Voronoi's cells associated to this element. The Voronoi' cell is a convex polygon containing all the points of the territory which are closer for the chosen element than to any other element (15). In our case, we use all the entities present on a molecular map for defining the Voronoi's cells. The polygons are precomputed for each map and post-processed to avoid too large polygons, and are colored according to the data values and the options defined by the user. In map staining, the colorful decorations of the pathway map are replaced by a black-and-white background in order to avoid mixing the colors. Using map staining allows to appreciate large-scale trends in molecular data mapped onto biological networks, when observed "from far".

RESULTS

We show that NaviCell Web Service can be used for simultaneous analysis of different types of high-throughput data in several case studies:

(1) comparing two prostate cancer cell lines using the cell cycle pathway map and transcriptomics and mutation data;

(2) using a large map of Atlas of Cancer Signalling Network (http://acsn.curie.fr (**?**) for visualizing ovary cancer data downloaded from The Cancer Genome Atlas (16) (shown in Supplementary Text ???);

(3) using the map of molecular interactions involved in Alzheimer disease (17) for visualizing the transcriptome data collected for different brain areas (18).

More examples of using NaviCell Web Service for data visualization can be found in NaviCell Web Service user guide and case studies provided at http://navicell.curie.fr/pages/nav_web_service.html

Comparing transcriptomes of prostate cancer cell lines

As an illustration of the use of maps in comparing two cancer genomic profiles, we selected two prostate cancer cell lines from the Cancer Cell Line Encyclopedia (19)): a prostate hormone-sensitive tumor cell line (LNCAP), and a prostate hormone-resistant tumor cell line (DU145). We gathered gene expression, copy number and mutation data from the two cell lines and mapped them onto the cell cycle map (20).

If we consider the cell line data as the mean expression of genes of a population of asynchronous cells, most LNCAP cells seem to be expressing genes from the early state G1 (Figure **??**A, upper-right area) and the G1-S checkpoint (Figure **??**A, lower-left area) while most DU145 cells express genes from the later stages of the cell cycle (Figure 2a, G1-late, S-phase and G2-phase areas). We notice that relatively few genes are mutated, amplified or lost in these two cell lines (Figure **??**A and **??**B, square and triangle glyphs).

By zooming on the map, we can determine that for the LNCAP cells, the most important gene alterations are: the amplification of DP1 (present in most complexes involving E2F1), and the homozygous loss of E2F2, transcription factor. The mutations concern mainly genes from the apoptotic pathway: ATR or CHEK2. As for the expression, the more noticeable trend is that the expression of the cell cycle inhibitors, such as RBL2, p21, CDC25C, is still high, whereas the expression of the G2 cyclin, for instance, CyclinA is low.

For the DU145 cells, genes involved in later stages of the cell cycle seem to be more expressed compared to the LNCAP cells. It can be that if the cells were arrested in LNCAP, they are more advanced in the cycle in DU145 cells by overpassing the G1/S checkpoint and are arrested at the spindle checkpoint. They are more prone to proliferate than the LNCAP cells. The expression of some cyclins seems to confirm this fact: CyclinB, CyclinD and CyclinH are higher than in LNCAP. Some means to stop the cycle seem to be kept though, with Cdc20 and Cdc25 expression high.

We verified **: AZ: This sounds like experimental validation using Q-PCR and misleading** the expression of KI-67, a marker of proliferation, and its expression is indeed higher in DU145 than in LNCAP cells, which tends to confirm our hypothesis that DU145 are more proliferative than LNCAP cells.

Creating molecular portrait of Alzheimer's disease

[Inna's contribution]

Table 2. Comparison of the NaviCell Web Service features with similar web sites for pathway-based data visualization.

Features	Na	Re	KE	iP	Bc	Pa
Map: navigation	●	●		●	●	●
Map: simple zooming	●	●	●	●	●	●
Map: semantic zooming	●					
Visualization: node coloring		●				●
Visualization: heatmaps	●				●	
Visualization: barplots	●				●	
Visualization: glyphs	●					
Visualization: map staining	●					
Data mapping: gene lists	●	●	●	●	●	●
Data mapping: expression data	●	●			●	●
Data mapping: copy-number data	●					
Data mapping: mutation data	●					
Data mapping: metabolomic data		●				
Data mapping: interactions		●				
Programmatic access: RESTful web	●	●	●			
Programmatic access: data visual.	●					

Abbreviations: Na: NaviCell Web Service, Re: Reactome(21), KE: KEGG(22), iP: iPath(23), Bc: BioCyc(24), Pa: PATIKAweb(25).

DISCUSSION

We have compared the features of the NaviCell Web Service for data visualization with similar tools. We selected a number of web-based tools that are providing similar functionalities, i.e. easy pathway browsing functions and interactive data visualization capabilities. For the comparison, we have focused on the features related to map navigation, molecular data types, graphical representations, and programmatic access. Table 2 display the results. NaviCell Web Service offers more features than other tools in terms of map navigation, for data visualization with additional graphical representations (such as map staining) and more data types, and finally extended support for programmatic access from different computer languages.

To extend the scope of NaviCell Web Service, we are currently working on porting some of the most used pathway databases into the CellDesigner format which can be used after for displaying in NaviCell, with a possibility of data visualization.

CONCLUSION

NaviCell Web Service should contribute to the growing set of highly demanded tools for molecular biology allowing visualization of "omics" data in the context of biological network maps.

ACKNOWLEDGEMENTS

Agilent. PIC SysBio. INVADE. COMET.

Conflict of interest statement. None declared.

REFERENCES

1. Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., and Gavin, A.-C. (Mar, 2010) Visualization of omics data for systems biology.. *Nat Methods,* **7**(3 Suppl), S56–S68.
2. Barillot, E., Calzone, L., Hupe, P., Vert, J.-P., and Zinovyev, A. (2012) Computational Systems Biology of Cancer, Chapman & Hall, CRC Mathematical & Computational Biology, .
3. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (Jan, 2012) KEGG for integration and interpretation of large-scale molecular data sets.. *Nucleic Acids Res,* **40**(Database issue), D109–D114.
4. Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., and D'Eustachio, P. (Jan, 2014) The Reactome pathway knowledgebase.. *Nucleic Acids Res,* **42**(Database issue), D472–D477.
5. Arakawa, K., Kono, N., Yamada, Y., Mori, H., and Tomita, M. (2005) KEGG-based pathway visualization tool for complex omics data.. *In Silico Biol,* **5**(4), 419–423.
6. van Iersel, M. P., Kelder, T., Pico, A. R., Hanspers, K., Coort, S., Conklin, B. R., and Evelo, C. (2008) Presenting and exploring biological pathways with PathVisio.. *BMC Bioinformatics,* **9**, 399.
7. Luo, W. and Brouwer, C. (Jul, 2013) Pathview: an R/Bioconductor package for pathway-based data integration and visualization.. *Bioinformatics,* **29**(14), 1830–1831.
8. Nishida, K., Ono, K., Kanaya, S., and Takahashi, K. (2014) KEGGscape: a Cytoscape app for pathway data integration.. *F1000Res,* **3**, 144.
9. Kuperstein, I., Cohen, D. P., Pook, S., Viara, E., Calzone, L., Barillot, E., and Zinovyev, A. (2013) NaviCell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. *BMC Systems Biology,* **7**(1), 100.
10. Fielding, R. T. and Taylor, R. N. (2002) Principled design of the modern Web architecture. *ACM Transactions on Internet Technology (TOIT),* **2**(2), 115–150.
11. Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K.-Y., and Kitano, H. (2011) Software for systems biology: from tools to integrated platforms. *Nature Reviews Genetics,* **12**(12), 821–832.
12. Chu, L., Scharf, E., and Kondo, T. (2001) GeneSpring: tools for analyzing microarray expression data. *Genome Info,* **12**, 227–229.
13. Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008) CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proceedings of the IEEE,* **96**(8), 1254–1265.
14. Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T. C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D. B., Sander, C., Sauro, H., Snoep, J. L., Kohn, K., and Kitano, H. (Aug, 2009) The Systems Biology Graphical Notation.. *Nat Biotechnol,* **27**(8), 735–741.
15. Aurenhammer, F. (1991) Voronoi diagrams: a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR),* **23**(3), 345–405.
. Kuperstein, I., Grieco, L., Cohen, D., Thieffry, D., Zinovyev, A., and E., B. (2015) The shortest path is not the one you know: application of biological network resources in precision oncology research. *Mutagenesis,* **In press**.
16. Network, C. G. A. R. (Jun, 2011) Integrated genomic analyses of ovarian carcinoma.. *Nature,* **474**(7353), 609–615.
17. Mizuno, S., Iijima, R., Ogishima, S., Kikuchi, M., Matsuoka, Y., Ghosh, S., Miyamoto, T., Miyashita, A., Kuwano, R., and Tanaka, H. (2012) AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease.. *BMC Syst Biol,* **6**, 52.
18. Hokama, M., Oka, S., Leon, J., Ninomiya, T., Honda, H., Sasaki, K., Iwaki, T., Ohara, T., Sasaki, T., LaFerla, F. M., Kiyohara, Y., and Nakabeppu, Y. (Sep, 2014) Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study.. *Cereb Cortex,* **24**(9), 2476–2488.
19. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling

of anticancer drug sensitivity. *Nature,* **483**(7391), 603–607.
20. Calzone, L., Gelay, A., Zinovyev, A., Radvanyi, F., and Barillot, E. (2008) A comprehensive modular map of molecular interactions in RB/E2F pathway. *Molecular Systems Biology,* **4**(1).
21. Croft, D., OKelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2010) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research,* p. gkq1018.
22. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research,* **28**(1), 27–30.
23. Letunic, I., Yamada, T., Kanehisa, M., and Bork, P. (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends in Biochemical Sciences,* **33**(3), 101–103.
24. Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., and López-Bigas, N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research,* **33**(19), 6083–6089.
25. Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetin-Atalay, R., and Ozturk, M. (2002) PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics,* **18**(7), 996–1003.

FIGURE LEGENDS

Figure 1. General architecture of the NaviCell Web service *server*. Client software (light blue layer) communicates with the server (red layer) through standard HTTP requests using the standard JSON format to encode data (RESTful web service, dark blue layer). A session (with a unique ID) is established between the server and the client browser (yellow layer) through Ajax communication channel to visualize the results of the commands send by the client. It is worth noticing that communication channels are bidirectional, i.e. the client software can send data (e.g. an expression data matrix) to the server, but it can also receive data from the server (e.g. a list of gene HUGO codes contained in a map).