Atelier Grands Graphes et Bioinformatique

Etienne E. Birmelé, Mohamed Elati, Blaise Hanczar, Lydia Boudjeloud-Assala

Atelier Grands Graphes et Bioinformatique

Organisateurs :
Etienne Birmelé (MAP5 - Université Paris Descartes)
Mohamed Elati (ISSB - Université d'Evry),
Blaise Hanczar (IBISC - Université Evry Val d'Essonne),
Lydia Boudjeloud-Assala (LITA EA 3097 - Université de Lorraine)

PRÉFACE

Le groupe de travail d'EGC *Grands Graphes et Bioinformatique* se réunit pour la deuxième fois dans le cadre des ateliers de la conférence EGC.

L'objectif de cet atelier est de réunir des chercheurs venant de plusieurs disciplines et s'intéressant au lien entre fouille de grands graphes et bioinformatique. De nombreux problèmes de biologie cellulaire reposent en effet sur des réseaux (similarité de séquences, interactions entre protéines, régulation entre gènes...) et les avancées technologiques actuelles rendent ces objets de plus en plus grands. L'étude de tels réseaux nécessite une étroite collaboration entre informaticiens, bioinformaticiens, mathématiciens et biologistes.

Les contributions de l'atelier 2016 illustrent en particulier deux grandes questions qui se posent actuellement à la communauté de bioinformatique au vu de l'afflux de données de type *omics*. La première est l'inférence du réseau d'intérêt, qu'il soit de type transcriptomique ou métabolique. Trois des contributions utilisent une telle étape avant de pousser leurs analyses et on peut voir qu'il n'y a pas consensus sur la manière d'aborder le problème. La seconde question au coeur de cet atelier est la manière d'extraire de la connaissance de ces réseaux, que ce soit en comparant différents réseaux entre eux ou en étant simplement capable de les visualiser de façon plus pertinente.

Nous tenons à remercier les auteurs pour la qualité de leurs contributions, ainsi que les comités de programme et d'organisation d'EGC qui ont permis à cet atelier de voir le jour.

Etienne Birmelé
Université Paris Descartes,
MAP5

Mohamed Elati
Université d'Evry,
ISSB

Blaise Hanczar
Université d'Evry,
IBISC

Lydia Boudjeloud-Assala
Université de Lorraine,
LITA EA 3097

TABLE DES MATIÈRES

Using local subnetworks to analyze genome-scale data

Frederik Gwinner [1,*], Gwénola Boulday [1,*], Claire Vandiedonck [2], Minh Arnould [1], Cécile Cardoso [1], Iryna Nikolayeva [4,5,6], Oriol Guitart-Pla [4], Elisabeth Tournier-Lasserve [1,3,†], Benno Schwikowski [4,†]

[1] Univ Paris Diderot, Sorbonne Paris Cité, INSERM UMR-S1161, F-75010 Paris, France
[2] Univ Paris Diderot, Sorbonne Paris Cité, INSERM U958, F-75010 Paris, France
[3] AP-HP, Groupe Hospitalier Saint-Louis Lariboisiere-Fernand-Widal, F-75010 Paris, France
[4] Systems Biology Lab, Institut Pasteur, F-75015 Paris, France
[5] Functional Genetics of Infectious Diseases Unit, Institut Pasteur, F-75015 Paris, France
[6] Univ Paris-Descartes, Sorbonne Paris Cité, F-75006 Paris, France

Most computational approaches for the analysis of genome-scale data in the context of interaction networks have very long running times, provide single or partial, often heuristic, solutions, and/or contain user-tuneable parameters.

We introduce local enrichment analysis (LEAN) for the identification of dysregulated subnetworks from genome-wide omics data sets. By substituting the common subnetwork model with a simpler *local* subnetwork model, LEAN allows efficient, exact, parameter-free, and exhaustive identification of local subnetworks that are statistically dysregulated, and directly implicates single genes for follow-up experiments.

Evaluation on simulated and biological data suggests that LEAN generally detects subnetworks better, and more clearly reflects biological similarity between experiments than standard approaches. A strong signal for the local subnetwork around Von Willebrand Factor (VWF), a gene which showed no change on the mRNA level, was identified by LEAN in transcriptome data in the context of the genetic disease Cerebral Cavernous Malformations (CCM). This signal was experimentally found to correspond to an unexpected strong cellular effect on the protein level of VWF. LEAN can be used to pinpoint statistically significant local subnetworks in any genome-scale data set.

Benno Schwikowski is a Research Director at Institut Pasteur, Paris, where he heads the Systems Biology Lab. Trained as a mathematician and computer scientist, he worked with Richard M. Karp at the University of Washington in Seattle, and started research in computational and network systems biology at the Institute for Systems Biology in Seattle. His research revolves around computational techniques for the discovery and modeling of complexity in genome-scale datasets, in particular around the human immune system.

DeDaL: Cytoscape 3 app for producing and morphing data-driven and structure-driven network layouts

Urszula Czerwinska[*,**,***] Laurence Calzone[*,**,***]
Emmanuel Barillot[*,**,***] Andrei Zinovyev[*,**,***]

[*]Institut Curie, 26 rue d'Ulm, Paris, France
http://bioinfo-out.curie.fr/projects/dedal/
[**]INSERM U900, Paris, France
[***]Mines Paris Tech, Fontainebleau, France

Abstract. DeDaL is a Cytoscape 3 app, which uses linear and non-linear algorithms of dimension reduction to produce data-driven network layouts based on multidimensional data (typically gene expression). DeDaL implements several data pre-processing and layout post-processing steps such as continuous morphing between two arbitrary network layouts and aligning one network layout with respect to another one by rotating and mirroring. The combination of all these functionalities facilitates the creation of insightful network layouts representing both structural network features and correlation patterns in multivariate data.

Availability. DeDaL is freely available for downloading at http:// bioinfo-out.curie.fr/projects/dedal/.

1 Introduction

One of the major challenges in systems biology is to combine in a meaningful way the large corpus of knowledge in molecular biology recapitulated in the form of large interaction networks together with high-throughput omics data produced at increasing rate, in order to advance our understanding of biology or pathology (Barillot *et al.*, 2012).

There exists numerous methods using the networks for making insightful high-throughput data analysis (Barillot *et al.*, 2012). These methods can be separated in three large groups, concentrating on: (1) mapping the data on top of a pre-defined biological network layout like in Kuperstein *et al.* (2013) and Shi *et al.* (2013), (2) identifying subnetworks from a global network possessing certain properties computed from the data (such as subnetworks enriched with differentially expressed genes) used by Ulitsky and Shamir (2007), Cline *et al.* (2007) and Alcaraz *et al.* (2012), and (3) using biological network structure for pre-processing the high-throughput data (for example, for "smoothing" the discrete mutation data) as demonstrated in Hofree *et al.* (2013).

Mathematically speaking, molecular entities exist in two metric spaces. The first one is the space of biological functions, where the distance between two molecules can be defined by

the number of steps (edges) in a graph defining pairwise functional relations (such as protein-protein interactions) along the shortest path connecting them. The other metric space is the data space, where the distance between two molecules is defined by the proximity of the corresponding numerical descriptors (such as expression profiles). The network distances are usually visualized by designing a 2D or 3D layout, representing the network structure. Visualization of distances in data space is achieved by data dimension reduction methods (such as PCA) projecting multidimensional vectors in 2D or 3D space.

We believe that in certain analyses, it could be insightful to construct the biological network layout based simultaneously on the network structure and the associated multidimensional data. One possible solution consists in applying data dimension reduction techniques. Unlike many other methods, the purpose of DeDaL is not to improve the visual appeal of the biological network layout, but to modify it in such a way that the trends in the associated data and exceptions from these trends would be detectable more easily.

2 Methods

2.1 Producing data-driven layout

Data-driven network layout (DDL) is produced by DeDaL by positioning the nodes of the network according to their projection from the multidimensional data space of associated numerical vectors into some 2D space. DeDaL implements three algorithms for performing this dimension reduction: (1) Projection onto a plane of several selected principal components; (2) Projection onto a non-linear surface approximating the multidimensional data distribution, i.e. principal manifold, computed by the method of elastic maps (Gorban and Zinovyev, 2010); (3) Using (1) or (2) preceded by network-based regularization (smoothing) of the data, based on computing the k first eigen vectors of the Laplacian matrix of the network graph and projecting data into this subspace (as suggested in Rapaport *et al.* (2007)).

DeDaL implements specific data pre-processing and resulting layout post-processing steps. Pre-processing steps include (1) selecting only nodes whose associated numerical vectors (imported as tables to Cytoscape) are sufficiently complete and (2) optional double centering of the data matrix. Post-processing of the resulting layout includes (1) avoiding overlap between node positions by moving them in a random direction at a small distance; (2) moving the outliers (nodes positioned too distantly from other nodes) closer to the barycenter of the data distribution; (3) placing nodes with missing data in the layout into the mean point of the position of their network neighbours. More detailed description of the methods and pre- and post-processing steps used in DeDaL is provided in Annexe 1.

2.2 Manipulating network layouts in DeDaL

In order to allow the comparison of the resulting DDLs with standard layouts produced by Cytoscape and to transform one into another, DeDaL implements simple layout morphing and aligning methods. Morphing of two network layouts is performed by a linear transformation, moving matched nodes along straight lines. DeDaL provides a convenient user dialog for morphing one layout into another so that the user can immediately appreciate the morphing result. The morphing operation provides poor results if one layout is systematically rotated or

flipped with respect to the node positions in another one. DeDaL allows aligning two network layouts by rotating, mirroring, and minimizing the Euclidean distance between two layouts.

3 Example of use

In order to illustrate the added value of DeDaL for visualizing expression data on top of relatively large networks, we constructed a tissue-specific subnetwork from HPRD global network of protein-protein interactions (PPIs) using the following approach. RNA-Seq data containing transcriptomes for 27 healthy human tissues were obtained from Fagerberg *et al.* (2014). In each profile, the genes were ranked according to their expression and the most significant largest connected component (LCC) of the global PPI network directly connecting the top ranked genes (OFTEN subnetwork) was identified using BiNoM plugin (Zinovyev *et al.*, 2008a), Bonnet *et al.* (2013b) (see detailed methodology description in Kairov *et al.* (2012)). After this step, the tissue-specific subnetworks that showed a significant score for the size of LCC were merged. This resulted in a network containing 1047 nodes, representing the top genes that are highly expressed in at least one tissue type, and 1986 edges representing direct PPIs between them.

DeDaL was applied to this network and the whole set of tissue transcriptomes. Double-centering and network smoothing with retaining only 5% of smallest eigenvectors was applied at the pre-processing step, and the non-linear principal manifold was computed for dimension reduction. Mapping transcriptomes of different tissues (spleen and brain) clearly highlights different network clusters with this layout (Fig. 1). The configuration of the clusters reflects the proximity of them in the data space of healthy tissue transcriptomes.

In order to objectively evaluate the advantage of using DeDaL for data visualization, we quantified how well the distances between the genes in the multidimensional space were reproduced on the 2D plane. We showed improved Pearson correlation between the distances from 0.1 (Force Directed layout) to 0.3 (DeDaL layout), leading to increase of correlation coefficient statistical significance by 18 orders of magnitude (Fig. 1, right bottom panel).

In Annexe 1 we also provide other examples of using DeDaL (transcriptome mapping, genetic interactions, visualizing attractors of Boolean model).

4 Conclusions

DeDaL Cytoscape plugin combines the classical and advanced data dimension reduction methods with the algorithms of network layouting inside Cytoscape environment. This ability can be used in a number of ways and for many applications, some of them are suggested in this paper.

The application of DeDaL is not limited to producing data-driven network layouts. More generally, DeDaL allows the application of dimension reduction of the multivariate data associated with the nodes of any Cytoscape network, optionally using the structure of the network, and exports the results for further analyses by any suitable algorithms.

5 Acknowledgements

We thank Eric Viara and Eric Bonnet for their help in implementing DeDaL and Loredana Martignetti for helping analyzing the data. All authors are members of the team "Computational Systems Biology of Cancer". The work is supported by ITMO Cancer SysBio program, (INVADE project) and, the grant "Projet Incitatif et Collaboratif: Computational Systems Biology Approach for Cancer" from Institut Curie and by Institut National de la Santé et de la Recherche Médicale (U900 budget).

References

Alcaraz, N., Friedrich, T., Kötzing, T., Krohmer, A., Müller, J., Pauling, J., and Baumbach, J. (2012). Efficient key pathway mining: combining networks and omics data. *Integrative Biology*, **4**(7), 756–764.

Barillot, E., Calzone, L., Hupe, P., Vert, J.-P., and Zinovyev, A. (2012). *Computational Systems Biology of Cancer*. Chapman & Hall, CRC Mathemtical and Computational Biology.

Bonnet, E., Calzone, L., Rovera, D., Stoll, G., Barillot, E., and Zinovyev, A. (2013a). Binom 2.0, a cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. *BMC Syst Biol*, **7**, 18.

Bonnet, E., Calzone, L., Rovera, D., Stoll, G., Barillot, E., and Zinovyev, A. (2013b). Practical use of binom: a biological network manager software. In *In Silico Systems Biology*, pages 127–146. Springer.

Bonnet, E., Calzone, L., Rovera, D., Stoll, G., Barillot, E., and Zinovyev, A. (2013c). Practical use of binom: a biological network manager software. *Methods Mol Biol*, **1021**, 127–146.

Calzone, L., Tournier, L., Fourquet, S., Thieffry, D., Zhivotovsky, B., Barillot, E., and Zinovyev, A. (2010). Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLoS Comput Biol*, **6**(3), e1000702.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., *et al.* (2007). Integration of biological networks and gene expression data using cytoscape. *Nature protocols*, **2**(10), 2366–2382.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis, B.-J., Shuteriqi, E., Tong, A. H. Y., van Dyk, N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B., Pál, C., Roth, F. P., Giaever, G., Nislow, C., Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A.-C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., and Boone, C. (2010). The genetic landscape of a cell. *Science*, **327**(5964), 425–431.

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., *et al.* (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics*, **13**(2), 397–406.

Gorban, A. N. and Zinovyev, A. (2009). Principal graphs and manifolds. *In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques, eds. Olivas E.S., Guererro J.D.M., Sober M.M., Benedito J.R.M., Lopes A.J.S.*

Gorban, A. N. and Zinovyev, A. (2010). Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *Int J Neural Syst*, **20**(3), 219–232.

Gorban, A. N., A., P., and Zinovyev, A. (2014). Vidaexpert: user-friendly tool for nonlinear visualization and analysis of multidimensional vectorial data. *Arxiv preprint*, (1406.5550).

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature methods*, **10**(11), 1108–1115.

Kairov, U., Karpenyuk, T., Ramanculov, E., and Zinovyev, A. (2012). Network analysis of gene lists for finding reproducible prognostic breast cancer gene signatures. *Bioinformation*, **8**(16), 773.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, **40**(Database issue), D109–D114.

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.

Moldovan, G.-L. and D'Andrea, A. D. (2009). How the fanconi anemia pathway guards the genome. *Annu Rev Genet*, **43**, 223–249.

Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T. K. B., Chandrika, K. N., Deshpande, N., Suresh, S., Rashmi, B. P., Shanker, K., Padma, N., Niranjan, V., Harsha, H. C., Talreja, N., Vrushabendra, B. M., Ramya, M. A., Yatish, A. J., Joy, M., Shivashankar, H. N., Kavitha, M. P., Menezes, M., Choudhury, D. R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C. K., Prasad, C. K., Kumar-Sinha, C., Deshpande, K. S., and Pandey, A. (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, **32**(Database issue), D497–D501.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J.-P. (2007). Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

Shi, Z., Wang, J., and Zhang, B. (2013). Netgestalt: integrating multidimensional omics data over biological networks. *Nature methods*, **10**(7), 597–598.

TCGA (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61–70.

Ulitsky, I. and Shamir, R. (2007). Identification of functional modules using network topology and high-throughput data. *BMC systems biology*, **1**(1), 8.

Zinovyev, A., Viara, E., Calzone, L., and Barillot, E. (2008a). Binom: a cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics*, **24**(6), 876–877.

Zinovyev, A., Viara, E., Calzone, L., and Barillot, E. (2008b). Binom: a cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics*, **24**(6), 876–877.

Annexe-1

A Details on the methods used by DeDaL

A.1 Double-centering the data matrix

The data matrix is optionally double-centered by subtracting from each matrix entry the mean value calculated over the corresponding matrix row and the mean value calculated over the matrix column, and by adding the global mean value computed over all matrix entries. This procedure allows to eliminate some global biases in the data such as the global differences in average fluorescence intensity of different probes in microarray data.

A.2 Network-based smoothing of data

Network data smoothing is made in DeDaL as it was suggested in Rapaport *et al.* (2007). For the given graph representing the network, its Laplacian and all its eigenvectors are computed. These vectors define a new orthonormal basis in the multidimensional data space. To smooth the values of the data matrix, the initial multidimensional vector associated to a data-point is projected into the subspace spanned by the first k eigenvectors of the graph's Laplacian. DeDaL smoothing parameter is the $p_{ns} = 1 - \frac{k-(n_c+2)}{N-(n_c+2)}, p_{ns} \in [0; 1]$, where n_c is the number of connected components in the graph and N is the number of nodes on the graph. Therefore, $p_{ns} = 0$ corresponds to $k = N$, i.e. when no smoothing is performed and all eigenvectors are used, while $p_{ns} = 1$ corresponds to $k = (n_c + 2)$ and first two non-degenerated eigenvectors are used to smooth the data (the data become effectively three-dimensional, with the first dimension corresponding to the average value of the data matrix computed over each connected component of the graph).

A.3 Exporting the pre-processed data

The results of pre-processing the data for a given network can be exported to a file. Actually, two files are created: one in a simple tab-delimited format suitable for further analysis in most statistical software packages and another file in the '.dat' format, suitable for analysis in ViDaExpert multidimensional data visualization tool (Gorban *et al.*, 2014). That way, network smoothing of an expression dataset can be done for further application in any machine learning algorithms (clustering, classification). For this purpose, DeDaL can be also used in a command line mode (see examples on the website, `http://bioinfo-out.curie.fr/projects/dedal/`).

A.4 Computing principal components

The principal components in DeDaL are computed using singular value decomposition, computed by the method allowing to use missing data values without pre-imputing them, as it is described in Gorban and Zinovyev (2009). Data points, containing more than 20% of missing values are filtered out from the analysis. DeDaL computes the 10 first principal components if there is more than 10 data points, and k principal components if there is $k + 1$ data points,

$k < 10$. After computing the principal components, DeDaL reports the amount of variance explained by each of the principal components.

A.5 Continuous layout morphing

Morphing two network layouts is performed by a simple linear transformation. A node having position (x_{11}, x_{12}) in the initial layout and the position (x_{21}, x_{22}) in the target layout is placed during the morphing procedure in the position $(p \times x_{21} + (1-p)x_{11}, p \times x_{22} + (1-p)x_{12})$, where $p \in [0; 1]$ is the morphing parameter representing the fracture of distance between the initial and target node positions along the straight line.

A.6 Aligning two network layouts by rotation and mirroring

Morphing between two network layouts might be meaningless if all nodes in one layout are systematically rotated or flipped with respect to the node positions in another layout. This situation is often the case when producing the pure data-driven layout and comparing it to the initial structure-driven layout. In this case, DeDaL allows minimizing the Euclidean distance between two layouts defined as the sum of squared Euclidean distances between all matched nodes with respect to all possible rotations and mirroring of one of the layouts. To do this, a user should simply check the corresponding checkbox in the user dialog before starting to apply layout morphing. Also, a user can align several network layouts to one chosen reference network layout, using a separate "Layout aligning" dialog. For example, it is usually useful to align the structure-driven layouts to the PCA-based data-driven layout.

A.7 Using DeDaL in command line mode

DeDaL can be used separately from the Cytoscape environment, in the command line mode, as it is explained on the DeDaL website with several examples. This is especially recommended for computing data-driven layouts for large networks containing more than ten thousand nodes. Command line mode allows applying all data pre-processing steps, including double-centering and network smoothing, saving the resulting network layout as a XGMML file and saving the eigenvector decomposition of the Laplacian of the network graph for future use.

B Implementation

DeDaL is implemented in Java and converted to a *simplified* Cytoscape 3 app. For computing linear and non-linear principal manifolds, DeDaL uses VDAOEngine Java library, developed by AZ (`http://bioinfo-out.curie.fr/projects/elmap/`). For computing the eigenvectors of a symmetric Laplacian matrix, Colt library has been used (`http://acs.lbl.gov/ACSSoftware/colt/`). Internal graph implementation is re-used from BiNoM Cytoscape plugin (Zinovyev *et al.*, 2008b; Bonnet *et al.*, 2013c,a). The source code of DeDaL is available at `http://bioinfo-out.curie.fr/projects/dedal`.

C Examples of use

C.1 Using The Cancer Genome Atlas transcriptome data and Human Protein Reference Database network

We used The Cancer Genome Atlas (TCGA) transcriptomic dataset for breast cancer (548 patients, (TCGA, 2012)) and Human Reference Protein Database (HRPD) database (Peri *et al.*, 2004) as a source of protein-protein interaction network.

Firstly, as an example of a small subnetwork, we selected proteins involved in Fanconi DNA repair pathway (Moldovan and D'Andrea, 2009) as it is defined in Atlas of Cancer Signaling Network (ACSN, `http://acsn.curie.fr`). For node coloring, we mapped the value of the t-test computed for the gene expression difference between the basal-like (one of the molecular subtypes of breast cancer, significantly contributing to the intertumoral variability) and non basal-like breast tumours. We have imported the TCGA data in Cytoscape and applied DeDaL for the transcription levels of the genes in the subnetwork (Figure 1).

One can see (Figure 2, top right) that the first principal component sorts the nodes accordingly to the t-test, because in this case the first principal component is associated with the basal-like breast cancer subtype. The second principal component gives additional information such as that the expression levels of BRCA2 and FANCE are differently modulated though both are upregulated in the basal-like subtype. Morphing the organic network layout with the PCA-based layout moves position of some of the genes, keeping the general pattern of PCA preserved, while better reflecting the network structure.

We have also applied PCA-based DDL to the subset of basal-like breast tumours (Figure 2, bottom left) which showed the specific role of BRCA1 gene in this subtype (which is known). Also, the position of USP1 gene has significantly changed with respect to the PCA-based DDL produced for the whole set of samples. This demonstrates the ability of DeDaL to produce network layouts specific for a particular cancer subtype.

Application of network smoothing is demonstrated at Figure 2, bottom middle. The layout preserves the general pattern of the PCA-based DDL, while better visualizing the network structure, and moving some proteins into a different position. For example, BRCA1 gene is moved to left because it is connected to several genes overexpressed in basal-like breast cancer subtype. Figure 2, bottom right, shows application of non-linear PCA to data dimension reduction. This network layout better resolves the relations between some gene expression levels such as FANCF and HES1 and the roles of BRCA1 and BRCA2 in Fanconi DNA repair pathway.

DDLs produced by DeDaL can serve to better visualize expression pattern in individual samples. Examples of using elastic map (elmap)-based DDL for distinguishing one randomly chosen basal-like and one non basal-like expression profiles of Fanconi pathway is shown in Figure 3. Unlike organic layout, DDL allows quickly evaluate the general trend of the expression profile and detect exceptions from this trend like USP1 gene, known to be a biomarker of genomic instability and Fanconi anemia phenotype, and overexpressed in both samples.

Secondly, we selected all proteins interacting with ESR1 protein (Figure 2). In this case, the second principal component shows, for example, that the expression levels of EGFR and CCNE1 are differently modulated though both are upregulated in the basal-like subtype. PCA layout also highlights a particular pattern of expression of some hub genes such as AR or EGFR, and shows that underexpressed genes in basal-like subtype forms more tightly con-

nected subnetwork. Morphing the original organic network layout with the PCA-based layout moves position of some of the proteins, keeping the general pattern of PCA preserved. For example, underexpressed PIK3R1, IGFR1 and ERBB2 genes are moved on the left because each of them is connected to several overexpressed genes. Application of network smoothing drives the hub genes to the center of the layout, because of averaging over the hub's neighbors. It produces more regular pattern of network connections but approximately conserves the neighborhood relations in PCA layout. Therefore, a combination of DeDaL methods allows different ways of mixing network structure and high-throughput data for producing new network layouts.

C.2 Visualizing genetic interactions

Genetic interactions between two genes happen in the case where their functions are synergistic (negative interactions) or mutually alleviating (positive interactions). The strength of genetic interactions is characterized by an epistatic score which quantifies deviation from a simple multiplicative model. In the global network of genetic interactions, each gene can be characterized by its epistatic profile, which is a vector of epistatic scores with all other genes (Costanzo *et al.*, 2010). It is shown that the genes with similar epistatic profiles tend to have similar cellular functions.

We applied DeDaL to create a DDL layout for a group of yeast genes involved in DNA repair and replication. The genetic interactions between these genes and the epistatic profiles (computed only with respect to this group of genes) were used from Costanzo *et al.* (2010). The definitions of DNA repair pathways were taken from KEGG database (Kanehisa *et al.*, 2012). Figure 5 shows the difference between application of the standard organic layout for this small network of genetic interactions and PCA-based DDL (computed here without applying data matrix double-centering to take into account tendencies of genes to interact with smaller or larger number of other genes). PCA-based DDL in this case groups the genes with respect to their epistatic profiles. Firstly, local hub genes RAD27 and POL32 have distinct position in this layout. Secondly, PCA-based DDL roughly groups the genes accordingly to the DNA repair pathway in which they are involved. For example, it shows that Non-homologous end joining DNA repair pathway is closer to Homologous recombination (HR) pathway than to the Mismatch repair pathway. It also underlines that some homologous recombination genes (such as RDH54) are characterized by a different pattern of genetic interactions than the 'core' HR genes RAD51, RAD52, RAD54, RAD55, RAD57,

C.3 Visualizing attractors of a Boolean model

In this example we used the Boolean model of cell fate decisions between survival, apoptosis and non-apoptotic cell death (such as necrosis) published in Calzone *et al.* (2010), to group the nodes of the influence diagram accordingly to their co-activation patterns in the logical steady states. The table of steady states was taken from Calzone *et al.* (2010) (Figure 6, top right) and used to compute the PCA-based DDL (Figure 6, bottom left). In this DDL, nodes in close positions have similar pattern of activation in steady states (such as RIP1 and RIP1K). We used morphing PCA-based DDL and the initial layout of the model (as it was designed in Calzone *et al.* (2010)) to visualize several stable states corresponding to different cell fates (Figure 7). In this layout co-activated nodes tend to form compact groups. Therefore,

DeDaL can be used to design layouts of mathematical models of biological networks, using the solutions of the model.

Scalability of DeDaL for large networks

DeDaL scales very well with respect to computing data projections into 2D space (see Figure 8). Even for the networks containing ten thousand nodes and more (such as the whole HPRD graph), DeDaL computes linear and non-linear data projections for few hundreds of samples in less than few tens of seconds on an ordinary laptop.

However, the network smoothing data pre-processing step implemented in DeDaL requires eigenvector decomposition of the Laplacian matrix of the network graph which scales in time as the third power of the number of nodes. While this computation remains relatively fast for relatively large networks (several minutes for a network of 2000 nodes, in our benchmark example), it drastically slows down when the size of the network grows above several thousands of nodes. In our benchmark example, eigenvector decomposition of the Laplacian of the whole HPRD PPI database required 7 hours on a regular laptop, which makes application of network smoothing data pre-processing not convenient for large networks. However, eigenvector decomposition of the Laplacian for a large graph can be done once and saved on the disk for future reuse. For example, on the DeDaL website we provide the pre-computed eigenvector decomposition for the Laplacian of the graph representing the whole HPRD database, and other decompositions for large PPI networks will be provided in the future. Using pre-computed eigenvector decomposition allows applying data network smoothing with large networks containing tens of thousands nodes in a reasonable time (few minutes).

Use of DeDaL with large networks containing tens of thousands of nodes is recommended in command line mode (see Implementation section). The computed network layout can be imported into Cytoscape environment and used for mapping high-throughput data on top of them.

Résumé

DeDaL est une application de Cytoscape 3 qui utilise des algorithmes de réduction de dimensions linéaires et non-linéaires, afin de créer des représentations de réseaux basées sur les données multidimensionnelles (par exemple des données d'expression de gènes). DeDaL intègre plusieurs étapes de préparation des données et de la représentation, tels qu'une transition continue entre deux représentations de réseaux arbitraires et l'alignement d'une représentation de réseau avec un autre par rotation et symétrie. La combinaison de ces fonctionnalités facilite la représentation pertinente de réseaux structurés et de corrélation dans des données multivariées.

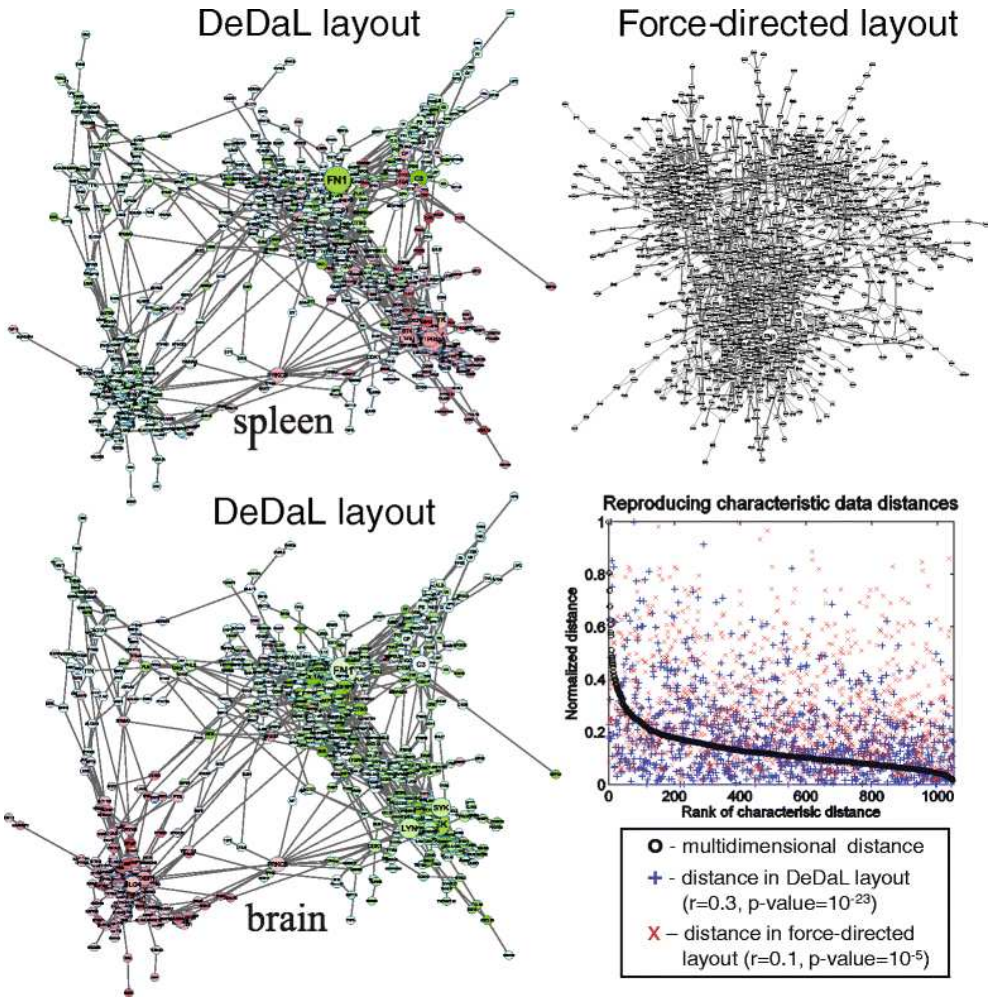

FIG. 1 – **Using DeDaL for visualizing the network and RNA-Seq expression data of tissue-specific genes.**. *RNA-Seq dataset for 27 healthy human tissues was used to defined a subnetwork of HPRD PPI database enriched in tissue-specific genes (see the text for explanations). Network smoothing followed by computation of principal manifold was applied to produce the data-driven network layout (DDL). Patterns of gene expression for two selected tissues (brain and spleen) are shown on top the constructed DDL, red color denotes higher expression, green color corresponds to lower expression. The sizes of the nodes are proportional to their connectivity degree in this network. On the left top panel application of the Force Directed layout is shown for comparison. On the left bottom panel results of quantitative comparison between multidimensional distance representation in DeDaL and Force Directed layout are shown. The most representative distances between the genes in the initial multidimensional space (see (Gorban and Zinovyev, 2010) for details) are ranked here from the largest to the smallest values*

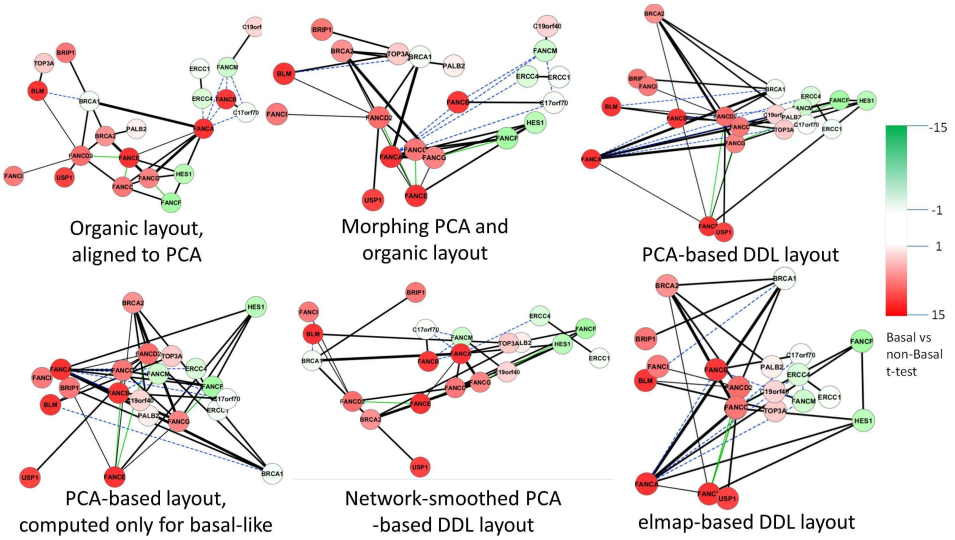

FIG. 2 – **Using DeDaL for visualizing Fanconi pathway in breast cancer**. *Top row from left to right: Standard organic layout, PCA-based DDL, morphing two previous layouts at half-distance. Bottom row from left to right: PCA-based DDL computed only for basal-like tumours (note change in position of BRCA1 gene), PCA applied to network-smoothed profile, DDL computed using elastic map (elmap) algorithm for computing non-linear principal manifold.*

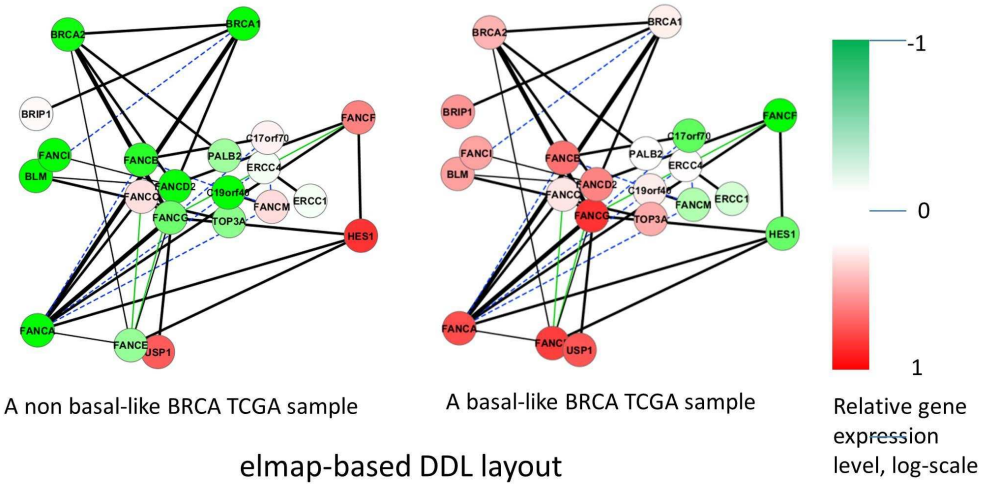

FIG. 3 – **Using DeDaL for showing individual sample gene expression profiles**. *Expression profiles on the Fanconi pathway genes for two randomly chosen samples (one basal-like and one non basal-like) from TCGA breast cancer cohorts are shown. The expression levels are computed as relative to the mean value over the whole cohort.*

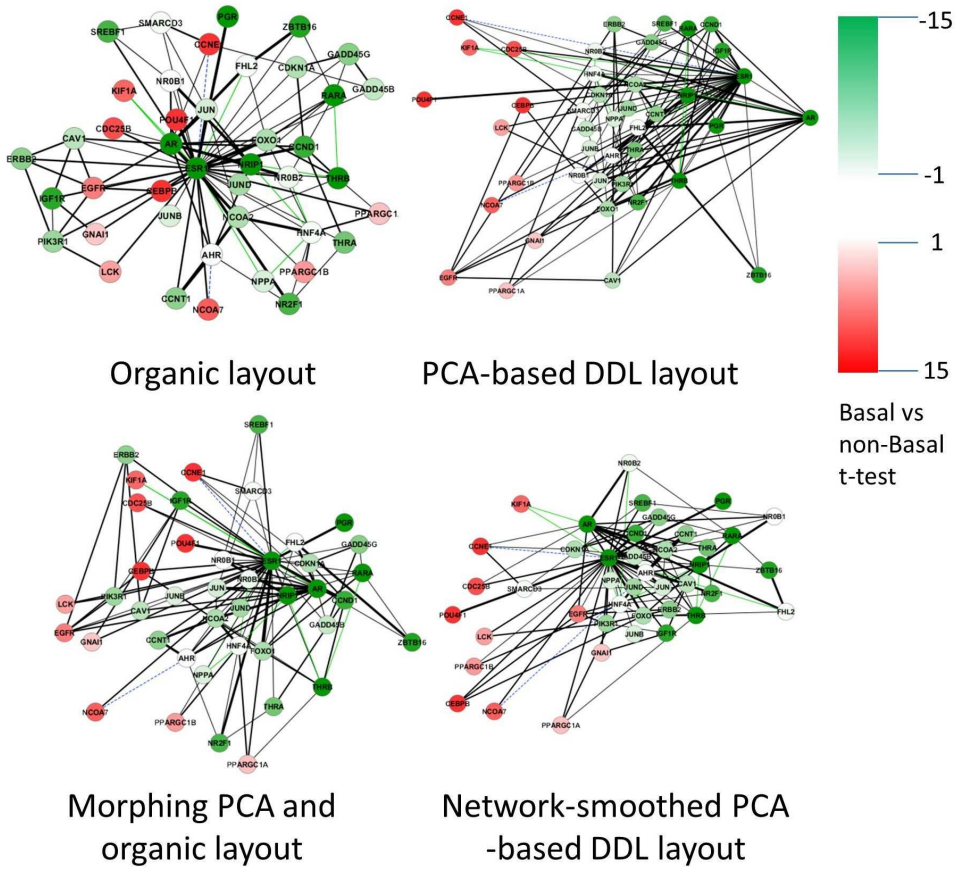

FIG. 4 – **Using DeDaL for visualizing network of genes interacting with ESR1**. *DeDaL allows mixing purely structure-driven network layout (top left) with purely data-driven network layout (top right) by morphing them (bottom left, which is the half-distance between two upper layouts). Bottom right is the same as PCA-based layout (top right) but network smoothing was performed before applying PCA.*

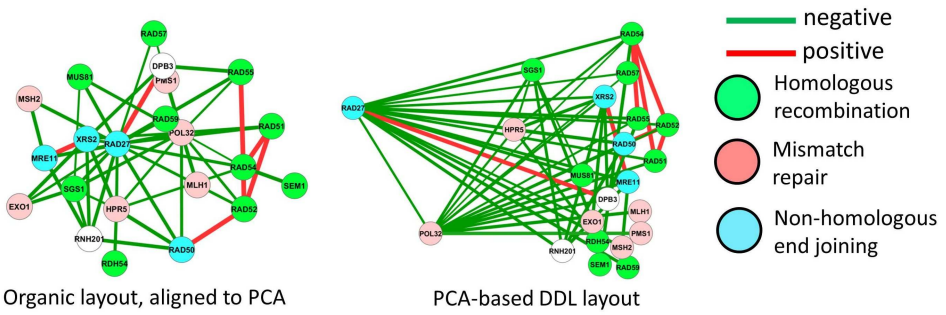

FIG. 5 – **Using DeDaL for visualizing network of genetic interactions between yeast genes involved in DNA repair**. *Red and green edges denote positive and negative genetic interactions correspondingly. Different node colors indicate three distinct DNA repair pathways in yeast.*

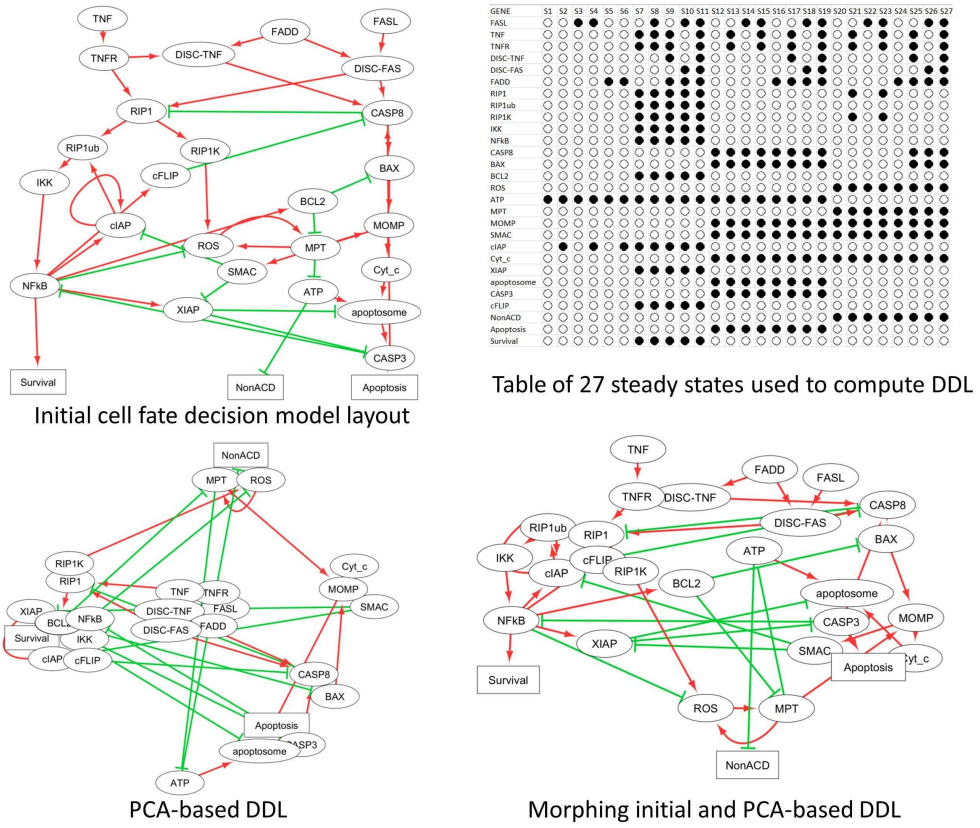

FIG. 6 – **Using DeDaL for visualizing results of a Boolean model simulation**. *Table of computed steady states is used to group the nodes with similar states in similar conditions (shown in top right corner). In the influence diagram green edges signify inhibitory and red edges - activating relations.*

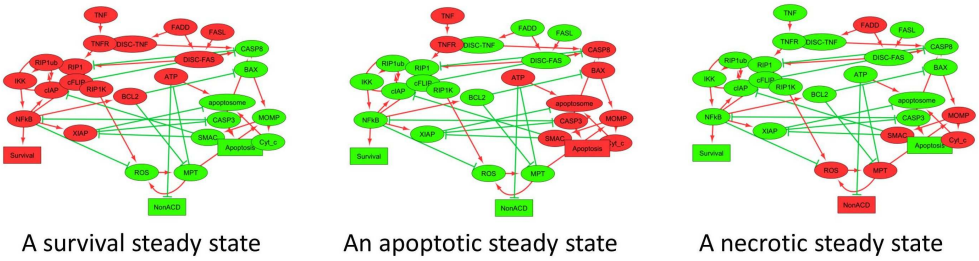

FIG. 7 – **Using DeDaL for visualizing results of a Boolean model simulation**. *Visualization of three steady states of the model, with green and red denoting inactive (FALSE) and active (TRUE) states of the node correspondingly.*

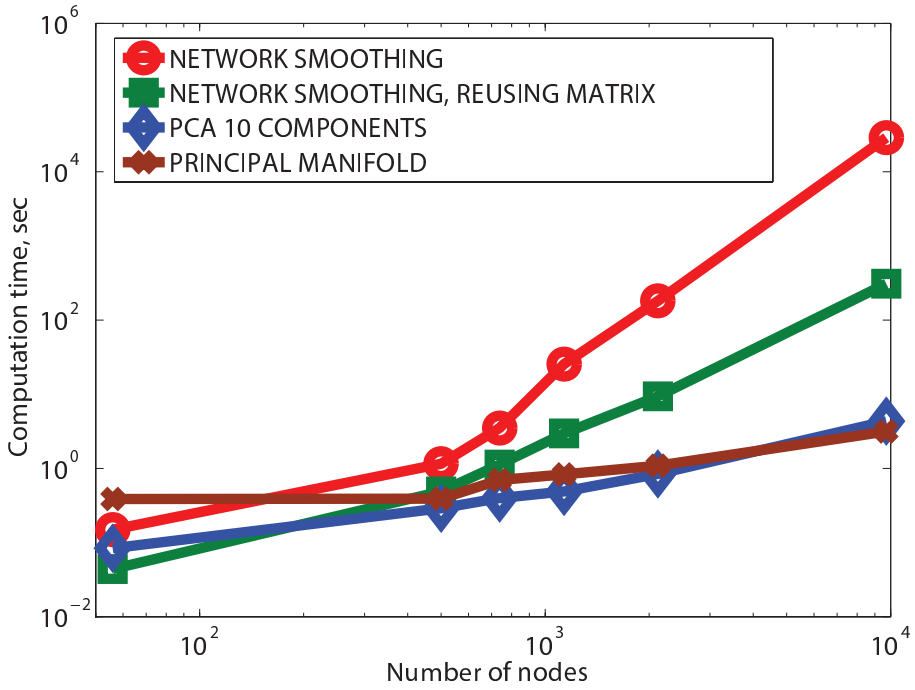

FIG. 8 – **Scalability of DeDaL for large networks**. *The figure shows the number of seconds needed for DeDaL to compute network smoothing (red line, circles), first ten principal components (blue line, rhombes) and the principal manifold (brown line, crosses) for a set of 100 ovarian cancer transcriptomes and a series of networks with increasing number of nodes (up to 10000 nodes in the whole HPRD PPI database). Network smoothing scaling is separately shown for the case of de-novo computation of the eigenvector decomposition of the network Laplacian (red line, circles) and for the case of using the pre-computed eigenvector decomposition (green line, squares). The benchmarking was done in the command line mode of DeDaL.*

Formalisation des réseaux biomoléculaires complexes

Ali Ayadi*, François de Beuvron* Cecilia Zanni-Merk*, Julie Thompson*

*ICube (UMR CNRS 7357) - 300 bd Sébastien Brant - BP 10413 - F-67412 Illkirch Cedex
{ali.ayadi,debeuvron,merk,thompson}@unistra.fr,

Résumé. La "transitabilité" d'un réseau biomoléculaire complexe exprime l'idée du pilotage de ce réseau d'un état non souhaité (en général associé à une maladie) à un état désiré. Dans ce article, nous présentons une formalisation du réseau biomoléculaire complexe et nous proposons une plate-forme basée sur les technologies sémantiques pour l'optimisation de la transitabilité de ces réseaux biomoléculaires.

1 Introduction

Dans le domaine de la biologie moléculaire, la conception et l'exploitation de nouvelles approches expérimentales s'intéressant au suivi de la structure des composants cellulaires et à leur construction ouvrent la voie à un nouveau domaine d'étude de la bioinformatique, la "Transitabilité". Selon Wu et al. (2014), *la transitabilité est l'orientation ou le pilotage d'un réseau biomoléculaire complexe d'un état non satisfaisant à un état désiré*.

Un réseau biomoléculaire est orchestré par les interactions de plusieurs molécules dans une cellule. Par défaut, une cellule vivante doit rester à un état normal (phénotype sain). Cependant, par une certaine perturbation inconnue ou stimulus, le réseau biomoléculaire peut changer d'un phénotype sain à un phénotype malade. Il est ainsi indispensable de trouver des mécanismes pour faire évoluer le réseau biomoléculaire du phénotype anormal au phénotype sain. Pour étudier les transitions de phénotypes, le réseau biomoléculaire est modélisé par un graphe dans lequel les molécules sont représentés par des nœuds et les interactions entre les molécules sont représentées par des arcs. En conséquence, les phénotypes cellulaires peuvent être définis par les états du réseau. Un état du réseau représente les expressions de toutes les molécules intervenant dans le réseau. Les changements phénotypiques ou les changements de comportement cellulaire seront donc décrits par une transition dynamique entre deux états du réseau Wu et al. (2014).

C'est dans ce contexte que nos travaux de thèse s'articulent autour du développement d'une plate-forme d'optimisation de la transitabilité des réseaux biomoléculaires complexes en proposant des mécanismes de pilotage des transitions de ces réseaux d'un état quelconque à un état spécifique. Nos travaux ont débuté en mai 2015, ayant comme objectif d'explorer le potentiel des techniques d'optimisation combinatoire et sémantiques pour optimiser et contrôler un réseau biomoléculaire complexe.

2 Modélisation du réseau biomoléculaire complexe

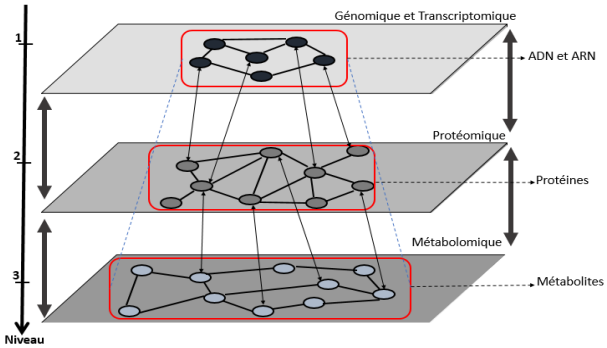

FIG. 1 – *Représentation multi-échelle du réseau biomoléculaire.*

La cellule est composée d'éléments moléculaires qui interagissent les uns avec les autres physiquement, fonctionnellement et logiquement formant ainsi un réseau biomoléculaire. Selon le type de ces éléments et de leurs interactions, nous distinguons trois types de réseaux qui ont fait l'objet de la plupart des études récentes Lombardi et Hörnquist (2007), Karp (2010) et Mones et al. (2012) (Figure 1) :

1. Le réseau génétique qui modélise les interactions entre à environ 21 000 gènes (ADNs et ARNs). Il est représenté par un graphe dirigé où les nœuds représentent les gènes et les arcs modélisent le type de régulation (activation ou inhibition) d'un gène sur un autre.
2. Le réseau protéine-protéine qui modélise les interactions ayant lieu entre environ 80 000 protéines. Ce réseau est représenté par un graphe non orienté où les nœuds sont les protéines et les arêtes non dirigées modélisent les liaisons entre ces nœuds. Ces types d'interactions dépendent de la fonction des protéines.
3. Le réseau métabolique modélise les réactions du métabolisme c'est à dire les réactions biochimiques entre les environ 42 000 métabolites réactants et produits. Un réseau métabolique est représenté par un graphe dont les nœuds sont des métabolites (réactants ou produits) et les arcs représentent le type de réaction biochimique transformant les réactants en des produits, elles sont marquées par le coefficient stœchiométrique du métabolite dans la réaction.

Ainsi, un réseau biomoléculaire peut être représenté par un graphe noté généralement par le quadruplet suivant :

$$G = (M, I, c, v)$$

Avec :

— M : l'ensemble des molécules formant le réseau représente les nœuds du graphe G : défini par un ensemble fini de sommets $M = \{m_1, m_2, \ldots, m_n\}$. Nous distinguons une partition tripartite de noeuds : les ensembles M_G des gènes (ADNs et ARNs), M_P

des protéines et M_M des métabolites. Nous pourrons ainsi écrire :

$$M = M_G \cup M_P \cup M_M$$

— I : l'ensemble des interactions entre les molécules formant le réseau. Représente les arcs du graphe G : défini par un ensemble fini d'arcs ou d'arêtes $I = \{I_1, I_2, \ldots, I_m\}$. L'arc $I = (m_i, m_j)$, (avec $m_i, m_j \in M$) qui part de m_i (origine) et arrive à m_j (destination) sera également noté $m_i \rightarrow m_j$. La partition des nœuds du graphe induit aussi une partition des arcs en divers types d'interactions :
  . trois interactions entre des composants moléculaires de même type : les interactions entre les gènes noté I_{GG} qui modélise le type de régulation entre les gènes (activation ou inhibition), I_{PP} qui modélise les associations stables ou transitoires entre les protéines et I_{MM} modélisant les interactions ayant lieu entre les métabolites (type de réaction chimique entre réactants et produits).
  . quatre interactions (parmi les 6 possibles) entre noeuds de réseaux différents : I_{GP} qui modélise les régulation des gènes sur les protéines, I_{PG} qui modélise l'influence des protéines sur les gènes à travers le facteur de transcription, I_{PM} représente les enzymes intervenant dans les réactions chimiques des métabolites (catalyse ou hydrolyse), I_{MP} modélise l'influence des métabolites sur les protéines.
  . deux interactions I_{GM} et I_{MG} ne sont pas prises en compte car il n'y a pas d'interaction directe entre les gènes et les métabolites.

$$I = I_{GG} \cup I_{PP} \cup I_{MM} \cup I_{GP} \cup I_{PM} \cup I_{MP} \cup I_{PG}$$

— c : la fonction qui associe à chaque nœud (molécule) du graphe la valeur de sa concentration et les deux seuils (minimal et maximal) qui la déclenche. Ces seuils que l'on note $S_{(M_i)_{min}}$ et $S_{(M_i)_{max}}$ représentent respectivement les limites de concentration minimale et maximale entre lesquels l'interaction suivante va se déclencher.
— v : la fonction qui associe à chaque arc une étiquette représentant la nature de l'interaction (hydrolyse, régulation positive, régulation négative, etc.).

L'état d'un réseau biomoléculaire est défini comme l'ensemble des valeurs représentant les niveaux de concentration dans ses noeuds à un instant t. L'état du réseau G à l'instant t est modélisé par le vecteur d'état suivant :

$$C_{(G)_t} = (c_1, c_2, \ldots, c_n)_t$$

où c_i (avec { i ∈ [1..n] }) est la valeur de concentration dans le nœud $m_i \in M$.

Une transition d'état dans le réseau peut se produire soit par un stimulus interne (augmentation de la concentration d'une molécule, par exemple) soit par un stimulus externe (la prise de médicament par exemple).

3 Suite de nos travaux

L'objectif général de cette thèse est de trouver un ensemble de stimulus externes optimal à appliquer pendant un intervalle de temps pré-fixé pour faire évoluer le réseau de son état

courant à un autre état souhaité. Notre approche innovante est basée sur l'utilisation conjointe des technologies sémantiques, d'optimisation combinatoire et de simulation.

Avec ce but, nos travaux futurs se poursuivront par le développement d'une plate-forme pour étudier les transitions des réseaux biomoléculaires d'un état quelconque à un état spécifique, basée sur trois modules :

1. Le module ontologique : Ce module utilise des technologies sémantiques pour générer de nouvelles connaissances inférées (la découverte de nouvelles associations sémantiques entre les molécules) afin d'affiner l'étude des transitions du réseau. Ce modèle prend en entrée l'ensemble des informations natives (état du réseau ses transitions sous forme de valeurs et de paramètres) introduites par l'expert et fournit au final le réseau inféré constitué de transition d'états natifs et inférés. Cet enrichissement par des méta-données et de nouvelles connaissances facilitera les prises de décision grâce à une gestion performante de la connaissance Zanni-Merk (2014).
2. Le module de simulation : Ce module reproduira au cours du temps le comportement dynamique de chaque composant du réseau. Ce simulateur adoptera le formalisme DEVS Discrete Event Specification de Zeigler et al. (2000).
3. Le module d'optimisation : Avec ce module, l'application des algorithmes d'optimisation combinatoire pour fournir un ensemble de séquences de transitions offrant le meilleur pilotage du réseau d'un état à un autre, tout en décrivant les changements de valeurs ayant lieu à chaque composant du réseau.

Références

Karp, G. (2010). *Biologie cellulaire et moléculaire : Concepts and experiments*. De Boeck Supérieur.

Lombardi, A. et M. Hörnquist (2007). Controllability analysis of networks. *Phys. Rev. E 75*, 056110.

Mones, E., L. Vicsek, et T. Vicsek (2012). Hierarchy measure for complex networks. *PLoS ONE 7*(3), e33799.

Wu, F.-X., L. Wu, J. Wang, J. Liu, et L. Chen (2014). Transittability of complex networks and its applications to regulatory biomolecular networks. *Scientific reports 4*.

Zanni-Merk, C. (2014). *Knowledge technologies for problem solving in Engineering*. Mémoire d'Habilitation À Diriger des Recherches. Université de Strasbourg.

Zeigler, B. P., H. Praehofer, et T. G. Kim (2000). *Theory of modeling and simulation : integrating discrete event and continuous complex dynamic systems*. Academic press.

Summary

The "transittability" of a complex molecular network expresses the idea of steering this network from an undesired state (usually associated wih illness) to a desired state. In this article, we present a formalization of complex biomolecular networks and we propose a platform based on semantic technologies for optimizing the transittability of these biomolecular networks.

Comparaison de réseaux de gènes pour explorer le rôle des transcrits anti-sens

Marc Legeay[*,**], Béatrice Duval[*]

[*]LERIA - Université d'Angers - UNAM, 2 bd Lavoisier, 49045 Angers FRANCE
{prénom}.{nom}@univ-angers.fr
[**]Institut de Recherche en Horticulture et Semences (IRHS),
UMR1345 INRA-Université d'Angers-AgroCampus Ouest, Centre Angers-Nantes,
42 rue Georges Morel - BP 60057, 49071 Beaucouzé FRANCE

Résumé. Un des problèmes clés en bioinformatique est de comprendre les mécanismes de régulation au sein d'une cellule. Notre travail concerne l'étude des réseaux de gènes chez le pommier, avec la particularité d'y intégrer les acteurs encore mal connus que sont les ARN anti-sens. Pour explorer l'impact des transcrits anti-sens, nous proposons ici la comparaison des deux réseaux obtenus par une méthode d'inférence très conservative. Nous pouvons ainsi étudier les interactions directes entre les gènes qui sont modifiées si l'on fait intervenir les transcrits anti-sens dans la méthode d'inférence. Un ensemble de motifs caractéristiques autour de ces modifications permet de révéler des ensembles d'acteurs sens et anti-sens intéressants.

1 Motivations et données

Afin de modéliser les mécanismes de régulation au sein d'une cellule, de nombreux travaux se sont intéressés à la construction de réseaux de gènes à partir de la mesure d'expression de leurs transcrits. Les transcrits anti-sens (AS) sont des molécules d'ARN endogènes dont la totalité ou une partie de leur séquence est complémentaire avec d'autres transcrits. Ainsi, l'ARN anti-sens peut s'hybrider avec l'ARNm ce qui entraîne une dégradation de l'ARN et donc l'inhibition du gène. Les différents mécanismes d'action des transcrits anti-sens ne sont pas encore complètement connus (Pelechano et Steinmetz, 2013), mais il apparaît que leur rôle a été largement sous-estimé jusqu'ici. Ainsi une étude récente sur le pommier a détecté une transcription anti-sens pour 65% des gènes exprimés, ce qui laisse envisager un fort potentiel de régulation par les anti-sens, notamment dans les processus de réponse à un stress. Nous nous intéressons donc à la construction de réseaux de gènes faisant intervenir ces nouveaux acteurs que sont les ARN anti-sens.

Les données de transcription dont nous disposons ont été obtenues grâce à une puce qui couvre l'ensemble des gènes codants prédits sur le pommier avec la particularité d'inclure, pour chaque locus identifié, une sonde sens et une sonde anti-sens, soit au total 126 022 sondes. Nous nous intéressons à un contexte biologique qui est celui de la maturation de la pomme pour lequel nous disposons de 22 échantillons mesurés au moment de la récolte, données H (pour Harvest) et 60 jours après la récolte, données 60DAH (60 Days After Harvest). Après

normalisation des données, nous avons identifié 931 sondes Sens (S) et 694 sondes anti-sens (AS) différentiellement exprimées avec 200 gènes ($S \cap AS$) pour lesquels à la fois le sens et l'anti-sens ont une expression significativement différente. Ces 1 625 gènes forment ce que nous appelons dans la suite *les gènes d'intérêt, Sens et Anti-Sens*.

Une étude fonctionnelle de cet ensemble de gènes d'intérêt a montré l'importance de prendre en compte les transcrits AS (Legeay et al., 2015). En effet, en comparant les termes de GO significativement représentés dans l'ensemble S à ceux significativement représentés dans $S \cup AS$, nous avons mis en évidence *les termes révélés par les anti-sens*, c'est-à-dire les fonctions biologiques qui ne seraient pas apparues dans cette analyse fonctionnelle sans la prise en compte des transcrits anti-sens. Une des fonctions les plus significatives ainsi révélée est `response to cold` (avec une p-valeur de 10^{-5}), or ce processus est clairement en jeu dans cette expérience puisque les pommes sont conservées au froid après la récolte.

2 Comparaison de Réseaux d'interactions

De nombreuses méthodes d'inférence de réseaux de gènes ont été proposées ces dernières années. C3NET (Conservative Causal Core Network) (Altay et Emmert-Streib, 2010) est une méthode qui s'appuie sur l'information mutuelle et une étape de maximisation pour ne retenir pour chaque gène g qu'un seul voisin, celui qui donne la plus forte information mutuelle avec g. Cette démarche permet donc de produire un « cœur de réseau » où seules les interactions directes entre deux gènes g_1 et g_2 apparaissent, les interactions indirectes résultant d'une interaction avec un troisième gène n'étant pas considérées.

Afin d'étudier le rôle des anti-sens dans les réseaux de régulation, nous proposons de comparer deux réseaux obtenus grâce à C3NET, le réseau R_S impliquant uniquement les acteurs sens de S et le réseau R_SAS impliquant les acteurs sens et anti-sens de $S \cup AS$. Notre comparaison a pour but d'identifier quelles interactions directes sont modifiées si on prend en compte les transcrits anti-sens [1]. Même si l'information mutuelle entre 2 gènes est une mesure symétrique, l'algorithme utilisé par C3NET retourne d'abord un graphe orienté où chaque nœud se connecte à au plus un voisin ; pour cela, on calcule l'information mutuelle entre toutes les paires de gènes et on seuille la matrice obtenue pour ne retenir que les valeurs significatives ; puis on ne garde dans chaque ligne non nulle de la matrice que la valeur maximum, ce qui donne le voisin retenu dans cette ligne.

Lorsque nous intégrons les acteurs anti-sens, nous nous intéressons plus particulièrement aux sens qui se connectent à un anti-sens, car cela correspond à une liaison directe du réseau R_S qui est remplacée par une liaison sens vers anti-sens dans R_SAS.

Pour faire apparaître ces modifications, nous construisons le graphe G en ajoutant les arcs de R_S au réseau R_SAS. Dans une visualisation réalisée sous Cytoscape (Shannon et al., 2003), nous colorons les arcs de G suivant leur appartenance aux 2 réseaux : un arc est vert s'il est présent uniquement dans R_SAS, rouge s'il est présent uniquement dans R_S et gris s'il est présent dans les deux. Avec ce code couleur, une liaison directe du réseau R_S qui est remplacée par une liaison sens vers anti-sens dans R_SAS se traduit par un nœud sens dont partent un arc rouge et un arc vert (Figure 1a). Le graphe G obtenu pour les données de l'expérience 60DAH est présenté dans la Figure 1c.

1. La méthode DC3NET (Altay et al., 2011) ne permet pas de comparer des réseaux possédant des acteurs différents.

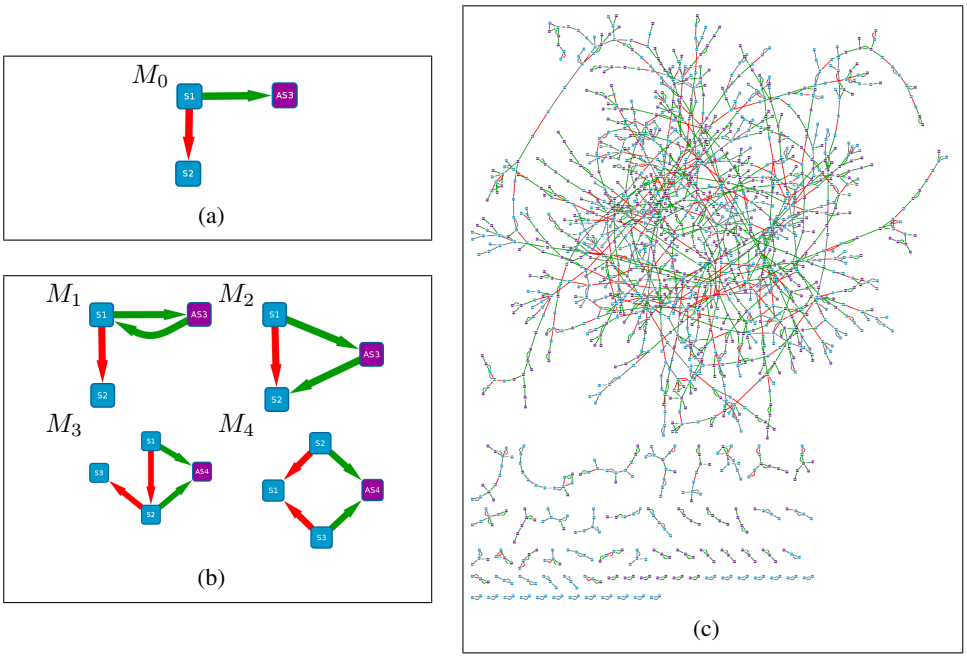

FIG. 1: Motifs et graphe de comparaison d'un réseau Sens avec un réseau Sens et Anti-Sens. Les nœuds bleus représentent les sens et les nœuds violets représentent les anti-sens. **(1a)** Schéma de modification d'une interaction directe. **(1b)** Motifs observés lorsqu'une interaction directe est modifiée. **(1c)** Graphe de comparaison entre les réseaux R_S et R_SAS pour l'expérience 60DAH.

Autour de ce schéma de modification d'interaction, nommé M_0, nous observons dans le graphe des motifs plus riches représentés dans la Figure 1b. Le motif M_1 dénote un lien fort entre un sens et un anti-sens. Le motif M_2 révèle une liaison entre $S1$ et $S2$ qui se trouve être indirecte car elle fait intervenir l'acteur intermédiaire $AS3$. Le motif M_3 représente une « attraction » de l'anti-sens : deux sens qui étaient connectés se lient maintenant au même anti-sens. On peut observer une variante de ce motif où $S3$ est confondu avec $S1$. Le motif M_4 décrit les changements de connexions : deux sens sont toujours reliés à un même acteur, mais cet acteur qui était un sens dans R_S est maintenant un anti-sens dans R_SAS. On peut observer une variante de ce motif où $S3$ se connecte à $AS4$. Ce motif permet d'identifier les anti-sens qui ont une information mutuelle importante avec des sens.

Nous avons construit les graphes de comparaison pour les expériences H et 60DAH. La Table 1 dénombre les motifs présents dans chacun de ces graphes. Parmi les 931 nœuds de R_S, environ 40% subissent une modification d'interaction, et ce dans les deux expériences.

D'autre part, nous avons croisé l'ensemble de gènes ainsi isolé avec l'étude fonctionnelle menée précédemment (cf Section 1). La fonction `response to cold` est représentée par 26 sens et 37 anti-sens. La Table 1 dénombre également les motifs contenant au moins un de ces gènes, ainsi que les gènes présents dans les motifs. En comparant l'expérience H avec 60DAH, on observe que près de la moitié des gènes `response to cold` interviennent dans un motif

Expérience		M_0	M_1	M_2	M_3	M_4
H	# motifs	378	117	26	31	8
	# motifs `response to cold`	31	0	6	1	3
	# gènes `response to cold`	28	0	7	1	3
60DAH	# motifs	373	102	21	16	5
	# motifs `response to cold`	29	1	2	1	7
	# gènes `response to cold`	24	1	2	1	8

TAB. 1: Nombre de motifs, nombre de motifs contenant au moins un gène `response to cold` et nombre de ces gènes apparaissant dans les motifs, sachant qu'il existe 63 gènes `response to cold` et 931 nœuds dans R_S.

M_0. Parmi eux, on observe 9 anti-sens qui sont présents dans M_0 à la fois dans H et 60DAH. De plus ces anti-sens ont la particularité d'avoir une expression qui diminue entre les deux expériences. L'idée ici est de combiner la comparaison des graphes avec l'étude fonctionnelle afin de soumettre à l'interprétation biologique des gènes ayant un comportement remarquable et associés à une fonction spécifique.

Références

Altay, G., M. Asim, F. Markowetz, et D. E. Neal (2011). Differential C3net reveals disease networks of direct physical interactions. *BMC Bioinformatics 12*(1), 296.

Altay, G. et F. Emmert-Streib (2010). Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology 4*(1), 132.

Legeay, M., B. Duval, J.-P. Renou, et J. Bourbeillon (2015). Construction et Analyse de Réseaux de Gènes Contextuels dans le Domaine Végétal. Poster JOBIM, 6–9 Juillet 2015. `www.inra.fr/jobim2015/Media/Fichier/Posters-PDF/Post-036`.

Pelechano, V. et L. M. Steinmetz (2013). Gene regulation by antisense transcription. *Nature Reviews Genetics 14*(12), 880–893.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, et T. Ideker (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research 13*(11), 2498–2504.

Summary

Understanding the regulation mechanisms in a cell is a key issue in bioinformatics. We study gene networks from apple data while integrating anti-sense transcripts. In order to explore the role of anti-sense transcripts in gene networks, we propose to compare two networks computed by a conservative reverse engineering method. Thus we can explore which direct interactions are modified when anti-sense transcripts are considered by the inference method. Specific graph patterns involving those modifications reveal interesting sense and anti-sense actors.

Integrated metabolic-regulatory modelling for diauxic shift analysis in S. cerivisiae

Daniel Trejo*

*iSSB, University of Evry, Genopole, 91030 Evry Cedex, France
trejo@issb.genopole.fr

Résumé. The diauxic shift is a complex biological process in which S. cerivisiae, under glucose depletion conditions, uses ethanol as main growth source. This change involves a reorganisation in all levels, as such it is of great interest comprehend and model this phenomenon. Here we present a brief outline of the modelling framework developed that seeks to integrate prior knowledge about transcriptional regulation and metabolic pathways into a single model that will then be used by an automated platform for performing experiments.

1 Introduction

When yeast cells grow in liquid media their primary carbon source is glycolysis that metabolises glucose. When glucose becomes scarse yeast enters a second phase of growth, though at a lower rate, after this second phase the cell enters into quiescence Galdieri et al. (2010).

As part of the Adalab (2015) initiative, we aim at developing computational tools for automated modelling and experimentation for studying the diauxic shift phenomenon. The main objectives are to use current knowledge drawn from transcriptomics and metabolomics literature in an unified modelling framework.

2 Integrated metabolic regulatory model

Our main idea is to employ transcriptomic data as background knowledge for a dynamic metabolic reconstruction of growth behaviour. The transcriptomic knowledge of 247 microarray assay for 5520 probes is collated in the Many Microbe Microarrays database (Faith et al., 2008) . We then analyse the data and reconstruct a co-regulatory network using the CoRegnet (Nicolle et al., 2015) utility.

CoRegnet allows not only to reconstruct a network of co-activators and co-regulators for a given set of transcriptomic data. It also performs influence analysis which helps identify transcriptional programs and is more robust to noise than traditional network reconstruction methods (Nicolle et al., 2015).

Having this background knowledge assembled we build a basic dynamic flux balance analysis solution using state of the art models of the whole yeast metabolic network, for example the model from Mo et al. (2007). This allow us to obtain a steady state approximation to growth

at each time step. By adjusting the constrains of this fba solution using the transcriptional information we can integrate both data sources in a single simulation. Then growth behaviour can be matched to observed growth curves and the plausibility of the results analysed.

3 Conclusions

We presented a brief modelling framework that aims at integrating metabolic and transcriptional regulation to study a single phenomenon. We employ state of the art methods in network reconstruction along with robust models of yeast metabolism. The main drawback being that we rely on a steady state approximation to dynamical problem. The main aim for current and future research is to adapt the framework into a more dynamical setting.

Références

Adalab (2015). AdaLab (an Adaptive Automated Scientific Laboratory). `http://www.adalab.mib.manchester.ac.uk/`.

Faith, J. J., M. E. Driscoll, V. A. Fusaro, E. J. Cosgrove, B. Hayete, F. S. Juhn, S. J. Schneider, et T. S. Gardner (2008). Many microbe microarrays database : uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic acids research 36*(suppl 1), D866–D870.

Galdieri, L., S. Mehrotra, S. Yu, et A. Vancura (2010). Transcriptional regulation in yeast during diauxic shift and stationary phase. *Omics : a journal of integrative biology 14*(6), 629–638.

Mo, M. L., M. J. Herrgård, G. Hannum, et B. Ø. Palsson (2007). Connecting extracellular metabolomic profiles to intracellular metabolic states in yeast.

Nicolle, R., F. Radvanyi, et M. Elati (2015). Coregnet : reconstruction and integrated analysis of co-regulatory networks. *Bioinformatics*.

Summary

The diauxic shift is a complex biological process in which S. cerivisiae, under glucose depletion conditions, uses ethanol as main growth source. This change involves a reorganisation in all levels, as such it is of great interest comprehend and model this phenomenon. Here we present a brief outline of the modelling framework developed that seeks to integrate prior knowledge about transcriptional regulation and metabolic pathways into a single model that will then be used by an automated platform for performing experiments.

Index