BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 15 2013, pages 1872–1878
doi:10.1093/bioinformatics/btt319

Systems biology

Advance Access publication June 7, 2013

Network2Canvas: network visualization on a canvas with enrichment analysis

Christopher M. Tan, Edward Y. Chen, Ruth Dannenfelser, Neil R. Clark and Avi Ma'ayan*

Department of Pharmacology and Systems Therapeutics, Systems Biology Center New York, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Networks are vital to computational systems biology research, but visualizing them is a challenge. For networks larger than ~100 nodes and ~200 links, ball-and-stick diagrams fail to convey much information. To address this, we developed Network2Canvas (N2C), a web application that provides an alternative way to view networks. N2C visualizes networks by placing nodes on a square toroidal canvas. The network nodes are clustered on the canvas using simulated annealing to maximize local connections where a node's brightness is made proportional to its local fitness. The interactive canvas is implemented in HyperText Markup Language (HTML)5 with the JavaScript library Data-Driven Documents (D3). We applied N2C to visualize 30 canvases made from human and mouse gene-set libraries and 6 canvases made from the Food and Drug Administration (FDA)-approved drug-set libraries. Given lists of genes or drugs, enriched terms are highlighted on the canvases, and their degree of clustering is computed. Because N2C produces visual patterns of enriched terms on canvases, a trained eye can detect signatures instantly. In summary, N2C provides a new flexible method to visualize large networks and can be used to perform and visualize gene-set and drug-set enrichment analyses.

Availability: N2C is freely available at http://www.maayanlab.net/N2C and is open source.

Contact: avi.maayan@mssm.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 24, 2013; revised on May 27, 2013; accepted on May 28, 2013

1 INTRODUCTION

Since the late 1990s, networks have been popularized for helping to increase our understanding of complex systems across various scientific disciplines (Barabasi and Albert, 1999; Watts and Strogatz, 1998), including systems biology and systems pharmacology. Commonly, networks are used to model complex systems by representing the various types of entities as nodes and their relationships as links. Network representation is useful for organizing the accumulating data about the system, which allows identifying novel non-trivial relationships. For example, network representation is useful for finding clusters in protein–protein interaction networks (Bader and Hogue, 2003) or for detecting network motifs (Milo *et al.*, 2002), such as feedforward and feedback loops. In addition, selected nodes with common properties, such as disease genes or differentially expressed genes, can be projected onto networks to highlight specific regions in the system, whereas forming a network can be a first step toward modeling the system's dynamics. In the fields of systems biology and systems pharmacology, networks are used to represent protein–protein interactions (Ito *et al.*, 2001), cell signaling pathways (Ma'ayan *et al.*, 2005), gene regulatory networks (Salgado *et al.*, 2013), metabolic networks (Jeong *et al.*, 2000), drug-target networks (Ma'ayan *et al.*, 2007) and other types of networks where nodes and links may represent different types of relationships (Lee *et al.*, 2004). Although network representation has been proven to be useful, large networks are difficult to visualize. The most popular approach to visualize networks is ball-and-stick diagrams. Many software tools are available to convert networks saved in a tabular format into a graphic ball-and-stick representation (Bastian *et al.*, 2009; Batagelj and Mrvar, 1998; Breitkreutz *et al.*, 2003; Dogrusoz *et al.*, 2006; Ellson *et al.*, 2002; Hu *et al.*, 2004; Shannon *et al.*, 2003; Wiese *et al.*, 2004). However, when there are more than ~100 nodes and 200 links, a situation that is common in systems biology and systems pharmacology, ball-and-stick diagrams fail to convey much useful information. Alternative methods such as TreeMaps (Shneiderman and Wattenberg, 2001), chord diagrams (Kassel and Turaev, 1995) or clustered adjacency matrix heatmaps can also visualize networks. These approaches have their own advantages and disadvantages, but most of these alternative methods have not been widely adopted. In the past, we have developed the software tool grid analysis of time-series expression (GATE), which is a tool to visualize time series gene expression data on a hexagonal grid movie (MacArthur *et al.*, 2010). Neighboring genes on the GATE hexagonal grid can be considered connected in a network where links between genes are established based on a gene–gene time-series correlation. This gave us the idea that any network can be visualized on a hexagonal grid.

More recently, we developed an algorithm called Sets2Networks to construct functional association gene–gene and drug–drug networks based on their shared properties (Clark *et al.*, 2012). Given gene- or drug-set libraries, typically used for gene-set enrichment analyses (Sherman and Lempicki, 2009; Subramanian *et al.*, 2005), the algorithm converts these gene- or drug-set libraries to functional association networks (FANs) based on various measures of gene or drug co-occurrence. In the FANs, genes are connected if they share many functional terms, such as belonging to the same pathway or regulated by the same transcription factor. Using the same approach, gene-set

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

libraries can be transposed—making the genes the terms, and the terms the set members associated with each gene; for example, a transposed gene-set library created from the Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathways database (Ogata *et al*., 1999) can be created where a gene/protein is the term, and members of each set are pathways that contain the gene/protein. Applying the Sets2Networks algorithm on transposed gene-set libraries results in term–term FANs. Similar networks have been developed for the Cytoscape (Shannon *et al*., 2003), plug-ins BiNGO (Maere *et al*., 2005) or EnrichmentMap (Merico *et al*., 2010), which display ball-and-stick diagrams of related gene ontology (GO) terms (Ashburner *et al*., 2000). With the popular tools BiNGO and EnrichmentMap, enriched terms are visualized on the network by highlighting the enriched nodes (GO terms) in different colors and size given input gene lists entered by users.

Here we implemented a new approach to visualize large networks. We project the network nodes onto a square grid canvas. The method and tool are called Network2Canvas (N2C). N2C arranges nodes on the canvas to maximize local connections where a node's brightness is made proportional to its local fitness. In addition, we use this network visualizing approach to perform gene-set enrichment analyses (Sherman and Lempicki, 2009; Subramanian *et al*., 2005). Conceptually, this approach provides similar functionality as BiNGO and EnrichmentMap, highlighting enriched terms on FANs. However, instead of visualizing term–term networks using balls-and-sticks, we visualize the networks of terms on a canvas where each square on the canvas represents a functional term, and terms are organized on the canvas based on their gene-set content similarity. This provides a condensed and appealing view of the results. Besides creating canvases for the three GO ontologies—biological process, cellular component and molecular function—as implemented for BiNGO and EnrichmentMap, we created canvases for 27 additional gene-set libraries. To make the plots interactive, web-based, lightweight and publication ready, we implemented the application using the JavaScript library Data-Driven Documents (D3) (Bostock *et al*., 2011).

2 METHODS

Gene-set library canvases: N2C contains 30 canvases representing 30 gene-set libraries belonging to six broad categories: transcription, pathways, ontologies, drugs/diseases, cell lines and miscellaneous. Although some of the libraries are borrowed from other sources (Chen *et al*., 2012; Culhane *et al*., 2010; Lachmann and Ma'ayan, 2009; Liberzon *et al*., 2011), including some from our previous publications, new gene-set libraries were added, including a gene-set library created from the Encyclopedia of DNA Elements (ENCODE) project (Rosenbloom *et al*., 2012), the Epigenomics Roadmap (Bernstein *et al*., 2010), the Cancer Cell Line Encyclopedia (Barretina *et al*., 2012) and the human and mouse gene atlases. Details about the construction of each gene-set library with references can be found in the supporting text. To create canvases from such gene-set libraries, we first transposed the libraries where the genes are the set labels and the terms are the set members. We then applied the Sets2Networks algorithm (Clark *et al*., 2012) to the resultant term-set libraries to create networks that connect terms from each library based on term–term gene-set content similarity.

Drug-set library canvases: N2C also contains six canvases representing drug–drug similarity networks and six canvases representing drug-property/drug-property similarity networks. These 12 canvases are made from the same six drug-set libraries linking drugs to their shared properties or properties to their shared drugs. The drugs in the six drug-set libraries are all approved by the Food and Drug Administration (FDA). Two libraries are created from side effect sources: The first side effect drug-set library is created from the Side Effect Resource (SIDER) side effect database (Kuhn *et al*., 2010), which has computational formatted information from drug insert labels. The second drug-set library is created from the FDA adverse events reporting system (FAERS) (Tatonetti *et al*., 2012). To create this library, we used 1 million recent FAERS reports. We first identified the top 40 most common severe side effects and then assigned side effects to drugs if the side effect was disproportionally associated with the side effect. We also created a drug-set library based on the drug first- and second-level Anatomical Therapeutic Chemical classification codes by processing information available from the DrugBank (Wishart *et al*., 2006). In addition, we also created a drug-set library based on the drug structure. Using the R library ChemmineR (Cao *et al*., 2008), the simplified molecular-input line-entry system (SMILES) strings of each FDA-approved drug were converted to a binary string representing various structural elements of the drugs. Then, to create a drug-set library, the structural elements were made the terms, and the drugs that contain these structural elements are members of the sets. The last two drug-set libraries are created from processing the Connectivity Map database (CMAP) (Lamb *et al*., 2006). We first processed the ranked files from CMAP to identify the top 100 up-and down-regulated genes for each experiment. Using these gene lists, we performed gene-list enrichment analysis using the KEGG pathway gene-set library (Ogata *et al*., 1999) or our own ChIP-seq enrichment analysis (ChEA) database/gene-set library (Lachmann *et al*., 2010). If the KEGG pathway or the ChEA transcription factor experiment was found to be enriched ($P<0.05$ after BH correction), then the drug was added to the pathway–drug-set library or the transcription factor drug-set library, respectively. To create networks from such drug-set libraries, we applied the Sets2Networks algorithm to the transposed and untransposed drug-set libraries to create six networks that connect FDA-approved drugs and six networks that connect terms that describe drug properties.

Converting networks to canvases: N2C creates canvases from any network, including FANs inferred from drug- or gene-set libraries. N2C contains two principal modules: the Annealer and the Visualizer. The Annealer converts an adjacency matrix representation of a network into a JavaScript Object Notation (JSON) file that contains the structure of the canvas; the Visualizer colors the canvas for display using D3 and outputs HTML and JavaScript files that can be viewed by modern browsers that support scalable vector graphics (SVG). To maximize the global fitness of the canvas, a Boltzmann probability function is used for the simulated annealing of the nodes on the canvas. Initially, nodes/terms are placed on the canvas randomly, and then pairs of nodes/terms are swapped repeatedly. As more time elapses, the probability of accepting a poor swap decreases (see supporting text for additional details).

Enrichment analysis and assessing clustering on the canvas: Enrichment analysis using N2C uses the standard method of the Fisher exact test applied on the overlap proportions between the query input list and the lists in the gene- or drug-set libraries. Resultant *P*-values are corrected using two alternative methods—Bonferroni and Benjamini–Hochberg—for multiple hypotheses testing. The enriched terms are highlighted on the canvas and are also output in a table with their corrected and uncorrected *P*-values. The enrichment analyses also produce *Z*-scores that show the degree of clustering of the enriched terms on the canvas. Subsets of nodes are considered significantly clustered when the position of the highlighted nodes on the canvas is significantly more clustered than what is expected for randomly placed nodes. To compute the degree of clustering, we first require a measure of average distance among the selected nodes: the mean nearest neighbor distance (Skellam, 1952). Because a random subset of nodes is distributed uniformly and randomly over the canvas, we also

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

need a measure of clustering relative to this random unclustered distribution (Clark and Evans, 1954). If we use the Manhattan distance, the probability of exactly zero nodes within a Manhattan distance of r of any given node is:

$$P(0, 2\rho r^2) = e^{-2\rho r^2} \tag{1}$$

where r is half the length of the squared canvas, and ρ is the number of selected squares per unit area. Next, we turn to the second event, which is that there is at least one node within the annulus defined by the radii r and $r+\delta r$. The area of this annulus is $4r\delta r$, so according to the Poisson distribution the probability of observing exactly zero events in this area is:

$$P(0, \rho(4r\delta r + 2\delta r^2)) = e^{-\rho(4r\delta r + 2\delta r^2)} \tag{2}$$

Repeating the steps as above, the probability distribution for the nearest Manhattan distance neighbors is:

$$p(r) = 4\rho r e^{-2\rho r^2} \tag{3}$$

This probability distribution can now be used to quantify the significance of any observed value of the mean nearest neighbor distance. For this, we need to know the expectation and variance under the null hypothesis of no clustering. This is calculated as follows:

$$E(r) = \int_0^\infty 4\rho r^2 e^{-2\rho r^2} dr = \sqrt{\frac{\pi}{8\rho}} \tag{4}$$

$$Var(r) = \int_0^\infty 4\rho r \left(r - \sqrt{\frac{\pi}{8\rho}}\right)^2 e^{-2\rho r^2} dr = \frac{4-\pi}{8\rho} \tag{5}$$

And the Manhattan distance z-score is given by:

$$z = \left(w - \sqrt{\frac{\pi}{8\rho}}\right)\sqrt{\frac{8N\rho}{\sqrt{4-\pi}}} \tag{6}$$

Note that here we retain the assumption of continuity. However, in our application the canvases are often not large enough and affected by the discrete nature of the squares. Running simulations for various size canvases, we noticed that power laws are a good approximation, so we fitted power laws to the expected distances between random nodes:

$$Mean = a_m \rho^{p_m} \tag{7}$$

$$\sigma = a_s n^{p_{s1}} w^{p_{s2}} \tag{8}$$

The fitted values are listed in Table 1. In the web application, we used the approximation and table to compute the level of clustering.

Network view and P-value view: N2C provides three views for each canvas. The default view shows the nodes in various shades of brightness, where node brightness is made proportional to its local fitness: the brighter a node is, the stronger its connections to its neighbors. The second view is called the *P*-value view. Here, only the top 20 enriched terms are highlighted where the brightness is based on the significance of the *P*-value. The third view is a network of enriched terms. Here, nodes are connected if they are enriched and close to each other on the canvas. All images generated by N2C are vector graphic SVG images that can be downloaded and imported into software tools such as Adobe Illustrator. The canvas view provides an interactive display of square labels, and clicking on squares provides more information about the terms and the set members.

Table 1. Fitted parameters for computing the clustering level of enriched terms

Variable	Estimate	Standard error	*t*-Statistic	*P*-value
a_m	0.62921	0.00571763	110.047	2.5×10^{-53}
p_m	−0.503301	0.00175724	−286.415	9.4×10^{-71}
a_s	0.328498	0.00617916	53.1622	1.8×10^{-30}
p_s1	−1.00728	0.00326769	−308.256	1.5×10^{-52}
p_s2	1.00939	0.004877	206.969	1.6×10^{-47}

3 RESULTS

3.1 Creating canvases from gene-set libraries and converting networks to a canvas

The workflow of N2C starts with a gene- or drug-set library coded as a gene matrix transposed (GMT) file or a network coded as an adjacency matrix. If starting with a drug- or a gene-set library, or other dataset that takes the form of a GMT file, the Sets2Networks algorithm converts the GMT file into an adjacency matrix. If starting with a network, the network is converted into an adjacency matrix. Python scripts with example data files and detailed explanations are provided on the N2C website help section. Using the adjacency matrix as input, the Annealer component of N2C organizes the nodes on a canvas, and once it is done annealing, it generates a JSON file that contains the information needed to display the canvas. The Visualizer script then creates the HTML and JavaScript files that are needed to display the interactive canvas in a browser (Fig. 1).

3.2 The N2C web application

We created 30 canvases from 30 different gene-set libraries (Fig. 2) and 12 canvases from 6 drug-set libraries (Fig. 3) and combined them into a web-based application available at

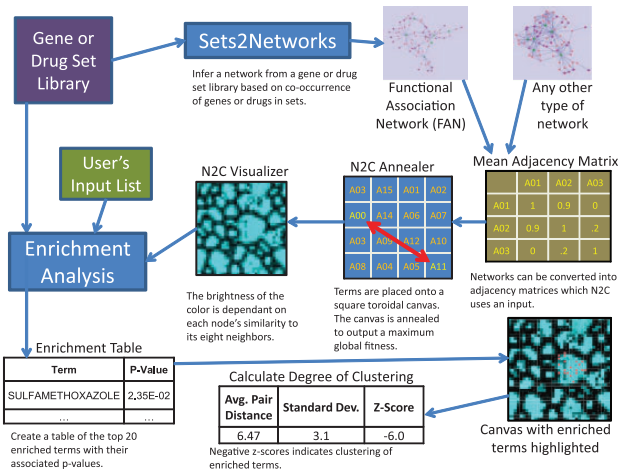

Fig. 1. The N2C workflow. Taking a mean adjacency matrix derived from a network as input, the Annealer places terms onto a square toroidal grid and anneals them for a user-defined duration, outputting a JSON file. The Visualizer reads this file and colors the canvas. Enrichment analysis highlights terms on the canvas, calculates the degree of clustering and outputs a table of the top 20 enriched terms and their *P*-values

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

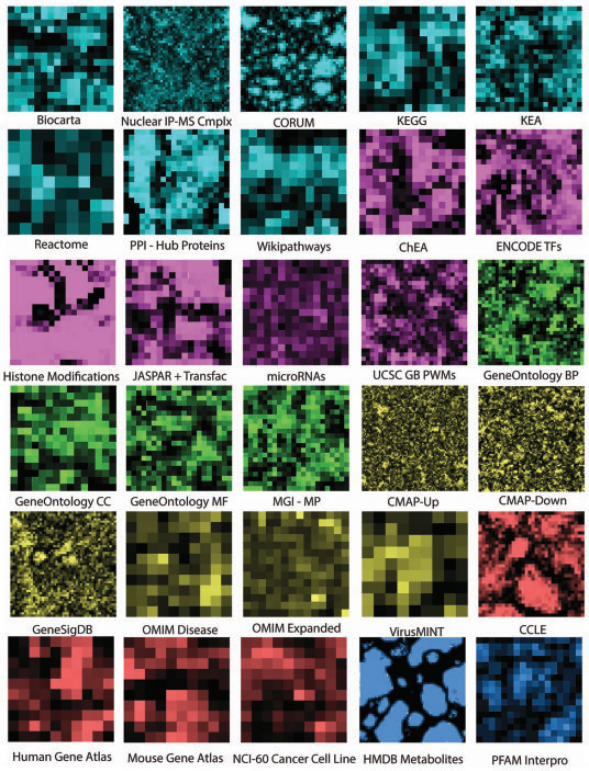

Fig. 2. Canvases for the 30 gene-set libraries. Each canvas represents a gene-set library. In each canvas, each node (square) represents a gene list associated with a gene-set library term. The terms are arranged on the canvas based on their gene-set content similarity. The brightness of each node is determined by its similarity to its eight neighbors. Canvases are color coded based on their category: pathways, cyan; transcription, purple; ontologies, green; disease/drugs, yellow; cell types, red and miscellaneous, blue

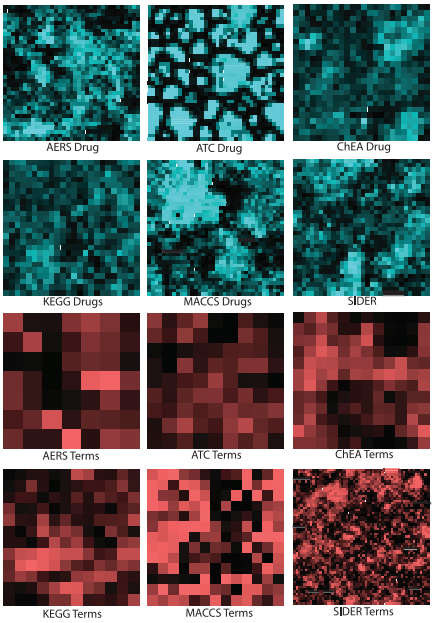

Fig. 3. Twelve canvases for the six drug-set libraries. Each canvas represents a drug-set library where each node (square) is an FDA-approved drug (blue canvases) or an FDA-approved drug property, e.g. a side effect or a structural element. Nodes are arranged based on their drug property similarity (for the drug canvases) or based on drug content similarity (for the drug property canvases). Nodes are colored based on their similarity to their neighbors: the brighter the spot, the higher the similarity. The six drug property set library canvases are the inverse of the six drug-set library canvases, all made from the original six drug-set library

http://www.maayanlab.net/N2C (Fig. 4). From this interface, users can perform gene- or drug-set enrichment analyses for data exploration and generate figures for publications. Later, we demonstrate how this can be accomplished in two case studies. In addition, Python scripts and a user manual are provided to enable users to create their own canvases from their networks.

3.3 Application of N2C to analyze gene expression data collected from embryonic stem cells

To demonstrate the usefulness of N2C, we performed enrichment analysis on lists of genes that are up-regulated after RNAi knockdown of various pluripotency transcription factors in mouse embryonic stem cells (mESCs) (Ivanova *et al.*, 2006). We first extracted lists of differential expressed up-regulated genes after knockdowns of the key well-studied factors in mESCs: Nanog, Oct4, Sox2, Esrrb and Tbx3. All these factors are critical for maintaining pluripotency because, when knocked down with short hairpin RNAs, mESCs differentiate into embryoid bodies after several days. Although it is known that these factors autoregulate each other and often physically interact (MacArthur *et al.*, 2009), not much is known regarding the

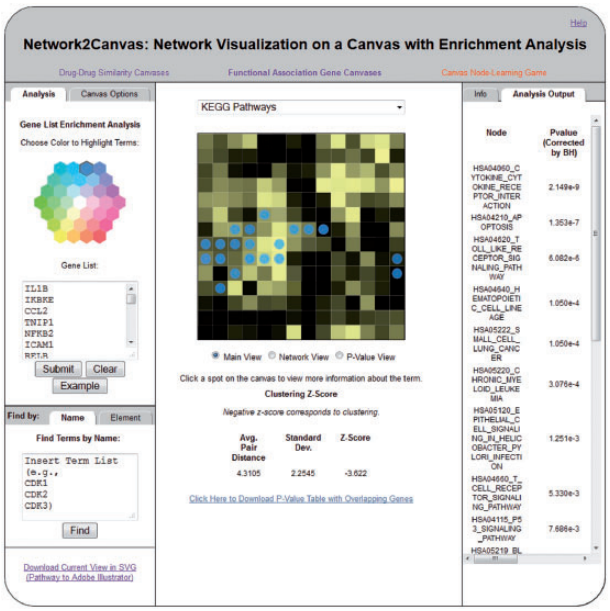

Fig. 4. Screenshot from the N2C web application. The interactive user interface of N2C provides users with the ability to select the desired canvas (KEGG is selected a default), change the canvas colors, highlight specific nodes on the canvas, perform enrichment analyses and explore the content of each node. The enrichment results are displayed in a table and can be exported

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

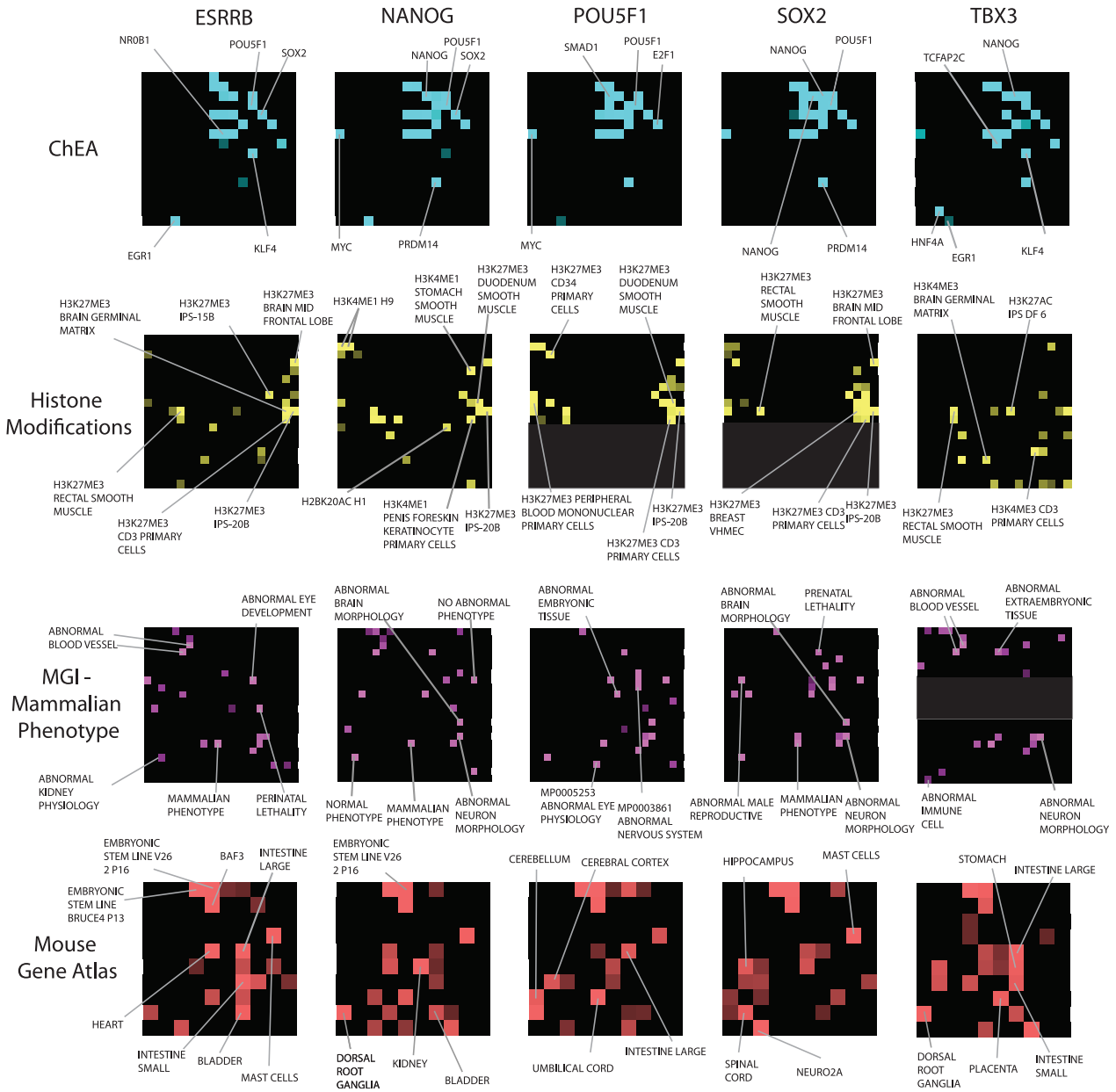

Fig. 5. Meta-signatures using various gene-set libraries applied to gene lists extracted from gene expression microarrays after knockdown of various transcription factors in mESCs. Only enriched terms are highlighted on the canvas. Brighter nodes represent lower *P*-value enriched terms

subtle differences of the effects of these knockdowns on lineage propensity. Using N2C, we generated enrichment analysis canvases resulting from ChEA, histone modifications, Mouse-Genome-Informatics/Mammalian-Phenotype (MGI-MP) and mouse cell atlas canvases. The results show that, overall, the various transcription factors induce similar patterns, but some differences are clearly visible (Fig. 5).

For example, Nanog, Esrrb and Tbx3 induce genes that are targeted by EGR1, but EGR1 is not enriched for Oct4 and Sox2 knockdowns. Esrrb is the only factor that does not induce the expression of genes regulated by Myc. Tbx3 up-regulates genes that are mostly regulated by the active histone marks H3K4me3, whereas the rest of the factors up-regulated genes are enriched for the repressive marks H3K27me3. Interestingly, Sox2 and Oct4 have a strong neuronal signature based on both MGI-MP and the mouse gene atlas, suggesting a propensity for differentiation toward neuroectoderm. It was previously described that Sox2 and Oct4 interact and regulate genes together (Chew *et al.*, 2005), whereas induction of Oct4 alone in neuronal progenitor cells is sufficient for induced pluripotency stem cell reprogramming (Kim *et al.*, 2009). Nanog, Tbx3 and Esrrb show suppression of kidney, intestine, bladder and endothelial cell phenotype genes. This suggests that the knockdown of these factors promotes the propensity toward mesoderm lineage differentiation. The advantage of the N2C visualization is that such similarities and differences can be seen immediately by the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 2. Two measures of clustering of enriched MACCS keys for severe side effects from FAERS on the drug structural elements canvas or computed from the drug structural elements adjacency matrix

Side effect	*z*-Score (canvas)	Average distance (direct)	*P*-value (direct)
Pancreatitis	−3.188	0.768	>0.0001
Coma	−3.157	0.772	0.002
Cardiac arrest	−3.018	0.676	0.0127
Neuropathy	−3.002	0.567	0.0311
Anxiety	−2.999	0.825	0.0003
Pneumonia	−2.768	0.579	0.0197
Depression	−2.648	0.893	0.0001
Bradycardia	−2.513	0.606	0.0232
Weight gain	−2.504	0.762	0.0036
Sleeplessness	−2.432	1.000	0.0003
Bronchospasm	−2.3	0.618	0.108
Hypoglycemia	−2.279	0.621	0.0034
Leukopenia	−2.279	0.632	0.0002

Note: The first column provides the *z*-score computed on the canvas as described in the method. The second column is the average distance between enriched terms on the drug structural element adjacency matrix. The third column is a *P*-value that examines the difference between the observed distance and the expected mean distance. The expected mean distance is computed as the average distance for the same number of, but randomly selected, structural elements from the drug structural element adjacency matrix.

human eye, condensing information into a colorful view of intuitive signatures.

3.4 Application of N2C to identify and visualize links between drug-induced side effects and chemical structure

Using the drug-set libraries created from FAERS, we ran the N2C enrichment analysis to identify the level of clustering of drug lists linked to 48 severe side effects on the drug chemical structure canvas. The hypothesis is that if drugs that are associated with a specific side effect are clustered on the drug-structure canvas, there might be a connection between drug structure and the side effect. Interestingly, for 18 of the 48 side effects, we observed significant clustering (z score>2.0) (Tables 2 and Supplementary Table S1, first column). This means that for 18 side effects there are common structural elements that are predictive of potential side effects. For the top three side effects—cardiac arrest, coma and pancreatitis—we display the drug canvas signatures (Fig. 6). The cardiac arrest and coma canvases share some structural elements, whereas the pancreatitis side effect canvas displays a completely different cluster of drugs with similar structures. Because the canvas visualization of the drug–drug association networks involves information loss due to the two-dimensional constraint and annealing process, we also examine the ability to associate side effects and structural elements directly from the adjacency matrices before the annealing. This was done to compare the level of information loss and to see whether the ranking of side effects is significantly altered. We computed the average network distance between structural elements for enriched side effects ($P<0.05$ without correction) to

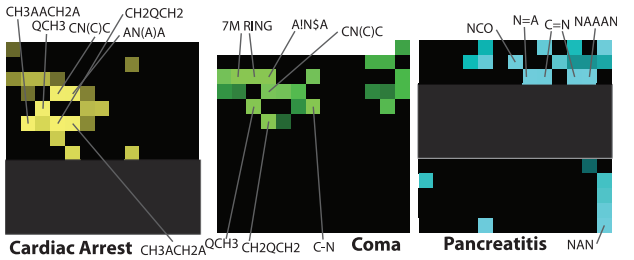

Fig. 6. Canvas for the 128 structural elements, also called Molecular ACCess System (MACCS) keys, using the *P*-value view in N2C show enrichment for MACCS keys for drug lists associated with the side effects of cardiac arrest, coma and pancreatitis

obtain a new ranked list of side effects that share related structural elements (Tables 2 and Supplementary Table S1, second and third columns). Although the two methods produce slightly different ranking of side effects, it is surprising that such a large majority of severe side effects can be predicted based on common structural elements by both methods. This may be important for *in silico* toxicity surveillance and for future drug design. Although quantitative structure–activity relationship models have been shown to mostly fail in the past in predicting most severe side effects, these initial drug enrichment results indicate that there may be useful information in approaching the problem in a slightly different angle as was done here. On the other hand, the methods used here to compute the relationship between structural elements and side effects can be improved in many ways, including better computation of expected distributions and better consideration of weights of enriched terms on the canvas.

4 DISCUSSION AND CONCLUSIONS

N2C provides an alternative method to visualize large networks and perform enrichment analyses. Although the approach is capable of condensing information, the projection of a high-dimensional object onto a two-dimensional canvas does cause significant information loss. This is compensated with the ability to better detect clusters from the network data by eye. However, it should be noted that the annealing approach may not reach the best possible fitness and may result in different layouts for the same input networks. The application of N2C to gene-set libraries currently only provides visualization of term–term libraries, but future applications can use the same approach to display all human or mouse genes. Seeing how groups of genes become active/inactive or expressed/silenced in different cell lines and conditions directly on such canvases may be illuminating. In addition, because each cell type has a unique expression signature and during development or disease there are transitions between cell types and cell phenotypes, movies of either enriched terms or clusters of genes can be created. The N2C approach and web-based tool generate visually appealing enrichment analysis signatures from sets of differentially expressed genes. Such signatures can be applied in high-throughput to datasets that profiled many conditions or many patients to detect patterns across large datasets (Duan *et al.*, 2013). Instead of visualizing the enriched terms for one list, the canvas can be used to display

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

the accumulation of enriched terms for many lists as a two-dimensional histogram.

The approach to link side effects to structural elements works for some drugs and some side effects but not for most drugs and side effects. A similar approach can be used to see if gene expression signatures induced by drugs can be predictive of side effects or if structural elements can be predictive of beneficial indications for various diseases. If such connections can be made, new drugs can be designed and tested by connecting structural elements to create novel drugs. The gene expression approach can be used to detect potential side effects before these manifest in patients.

Funding: This work is supported in part by NIH grants R01GM098316, R01DK088541, U54HG006097-02S1, P50GM071558 and the Irma T. Hirschl Career Scientist Award.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25.

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Barabasi,A.L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Barretina,J. *et al.* (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Bastian,M. *et al.* (2009) Gephi: an open source software for exploring and manipulating networks. In: *International AAAI conference on weblogs and social media*. Vol. 2, AAAI Press, Menlo Park, CA.

Batagelj,V. and Mrvar,A. (1998) Pajek-program for large network analysis. *Connections*, **21**, 47–57.

Bernstein,B.E. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.

Bostock,M. *et al.* (2011) D3: Data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, **17**, 2301–2309.

Breitkreutz,B.J. *et al.* (2003) Osprey: a network visualization system. *Genome Biol.*, **4**, R22.

Cao,Y. *et al.* (2008) ChemmineR: a compound mining framework for R. *Bioinformatics*, **24**, 1733–1734.

Chen,E.Y. *et al.* (2012) Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers. *Bioinformatics*, **28**, 105–111.

Chew,J.L. *et al.* (2005) Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol. Cell. Biol.*, **25**, 6031–6046.

Clark,N.R. *et al.* (2012) Sets2Networks: network inference from repeated observations of sets. *BMC Syst. Biol.*, **6**, 89.

Clark,P.J. and Evans,F.C. (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, **35**, 445–453.

Culhane,A.C. *et al.* (2010) GeneSigDB a curated database of gene expression signatures. *Nucleic Acids Res.*, **38**, D716–D725.

Dogrusoz,U. *et al.* (2006) PATIKAweb: a web interface for analyzing biological pathways through advanced querying and visualization. *Bioinformatics*, **22**, 374–375.

Duan,Q. *et al.* (2013) Meta-signatures identify two major subtypes of breast cancer. *CPT Pharmacometrics Syst. Pharmacol.*, **2**, e35.

Ellson,J. *et al.* (2002) Graphviz open source graph drawing tools. In: *Graph Drawing*. Springer, Berlin Heidelberg, pp. 594–597.

Hu,Z. *et al.* (2004) VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, **5**, 17.

International AAAI conference on weblogs and social media. Vol. 2. Menlo Park, CA: AAAI Press, 2009.

Ito,T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Ivanova,N. *et al.* (2006) Dissecting self-renewal in stem cells with RNA interference. *Nature*, **442**, 533–538.

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kassel,C. and Turaev,V. (1995) Chord diagram invariants of tangles and graphs. Institut de Recherche Mathématique avancée, Université Louis Pasteur et CNRS (URA 01).

Kim,J.B. *et al.* (2009) Direct reprogramming of human neural stem cells by OCT4. *Nature*, **461**, 649–653.

Kuhn,M. *et al.* (2010) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, **6**, 343.

Lachmann,A. and Ma'ayan,A. (2009) KEA: kinase enrichment analysis. *Bioinformatics*, **25**, 684–686.

Lachmann,A. *et al.* (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Lee,I. *et al.* (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.

Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Ma'ayan,A. *et al.* (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science*, **309**, 1078–1083.

Ma'ayan,A. *et al.* (2007) Network analysis of FDA approved drugs and their targets. *Mt. Sinai J. Med.*, **74**, 27–32.

MacArthur,B.D. *et al.* (2009) Systems biology of stem cell fate and cellular reprogramming. *Nat. Rev. Mol. Cell Biol.*, **10**, 672–681.

MacArthur,B.D. *et al.* (2010) GATE: software for the analysis and visualization of high-dimensional time series expression data. *Bioinformatics*, **26**, 143–144.

Maere,S. *et al.* (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Merico,D. *et al.* (2010) Enrichment Map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*, **5**, e13984.

Milo,R. *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Ogata,H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.

Rosenbloom,K.R. *et al.* (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.*, **40**, D912–D917.

Salgado,H. *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.*, **41**, D203–D213.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Shneiderman,B. and Wattenberg,M. (2001) Ordered treemap layouts. In: *Proceedings of the IEEE Symposium on Information Visualization 2001*. (Vol. 73078).

Skellam,J. (1952) Studies in statistical ecology: I. Spatial Pattern. *Biometrika*, 346–362.

Subramanian,A. *et al.* (2005) Gene-set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tatonetti,N.P. *et al.* (2012) A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *J. Am. Med. Inform. Assoc.*, **19**, 79–85.

Watts,D. and Strogatz,S. (1998) The small world problem, collective dynamics of small-world networks. *Nature*, **393**, 440–442.

Wiese,R., Eiglsperger,M. and Kaufmann,M. (2004) yfiles—visualization and automatic layout of graphs. In: *Graph Drawing Software*. Springer, Berlin Heidelberg, pp. 173–191.

Wishart,D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016