Bioinformatics, 38, 2022, i359–i368
https://doi.org/10.1093/bioinformatics/btac261
ISCB/ISMB 2022

DEMA: a distance-bounded energy-field minimization algorithm to model and layout biomolecular networks with quantitative features

Zhenyu Weng[1,†], Zongliang Yue[2,†], Yuesheng Zhu[1] and Jake Yue Chen[2,*]

[1]Communication and Information Security Lab, Institute of Big Data Technologies, Shenzhen Graduate School, Peking University, Shenzhen 518055, China and [2]Informatics Institute, School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Summary: In biology, graph layout algorithms can reveal comprehensive biological contexts by visually positioning graph nodes in their relevant neighborhoods. A layout software algorithm/engine commonly takes a set of nodes and edges and produces layout coordinates of nodes according to edge constraints. However, current layout engines normally do not consider node, edge or node-set properties during layout and only curate these properties after the layout is created. Here, we propose a new layout algorithm, distance-bounded energy-field minimization algorithm (DEMA), to natively consider various biological factors, i.e., the strength of gene-to-gene association, the gene's relative contribution weight and the functional groups of genes, to enhance the interpretation of complex network graphs. In DEMA, we introduce a parameterized energy model where nodes are repelled by the network topology and attracted by a few biological factors, i.e., interaction coefficient, effect coefficient and fold change of gene expression. We generalize these factors as gene weights, protein–protein interaction weights, gene-to-gene correlations and the gene set annotations—four parameterized functional properties used in DEMA. Moreover, DEMA considers further attraction/repulsion/grouping coefficient to enable different preferences in generating network views. Applying DEMA, we performed two case studies using genetic data in autism spectrum disorder and Alzheimer's disease, respectively, for gene candidate discovery. Furthermore, we implement our algorithm as a plugin to Cytoscape, an open-source software platform for visualizing networks; hence, it is convenient. Our software and demo can be freely accessed at http://discovery.informatics.uab.edu/dema.

Contact: jakechen@uab.edu

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

1 Introduction

Network analysis provides a fundamental way to understand the biomolecular mechanisms systematically and the dynamics of signaling events (Koutrouli *et al.*, 2020; Ma'ayan, 2011). When dealing with a huge, complex biological network, layout algorithms play a vital role in revealing the underlying structure and sub-structure or functional associations so that the critical components or modules can be captured easily. Ideal network layouts are expected to not only present a global view of items with more space-filling and fewer overlapping nodes but also highlight specific parts of the views coupling with biological annotations (Kobourov, 2012; Pinto *et al.*, 2010; You *et al.*, 2010). Therefore, the mined critical biomarkers and models can further assist disease diagnosis (Bock and Ortea, 2020; Drier *et al.*, 2013; Liu *et al.*, 2016; Livshits *et al.*, 2015; Pian *et al.*, 2021; Zhang and Chen, 2010), cancer subtyping (Lafferty *et al.*, 2021; Mallavarapu *et al.*, 2020; Zhang and Chen, 2013) and personalized medicine (Chen *et al.*, 2007; Hamburg and Collins, 2010; Raghavan *et al.*, 2017). Various analytical strategies have been applied to biological network analysis, such as sub-network analysis in MCODE (Bader and Hogue, 2003), hub gene identification and geneset enrichment analysis, to extract critical genomic information (Khan *et al.*, 2020). Many layout algorithms incorporate biological factors, such as hubs (highly connected nodes) (He and Zhang, 2006), into a weighted network to generate layouts. Compared with the traditional Fruchterman and Reingold (FR) layout model (Fruchterman and Reingold, 1991), the recently proposed Force-Directed (FD) algorithm, Organic Layout (OL) algorithm (Cline *et al.*, 2007), BioLayout (aka. edge-weighted FD layout) (Enright and Ouzounis, 2001) and Forcealtas2 (FA2) (Jacomy *et al.*, 2014) show a protein–protein interaction (PPI) network's hubs and overall network modularity clearly. To evaluate the PPI's strength in a PPI network, different statistical methods were developed to capture the PPI score (Chen *et al.*, 2017; Langfelder and Horvath, 2008; Schaefer *et al.*, 2012). PPI score is incorporated by edge-weighted FR (EWFR) provided by Cytoscape (Shannon *et al.*, 2003) to generate a layout where the length of the edges reflects the degree of the relation. In addition, since the gene-set enrichment analysis (Yue *et al.*, 2015) has become popular, Enrichment Map (EMap), a layout plugin for Cytoscape, is developed to present relationships among gene sets (Merico *et al.*, 2010). In EMap, nodes represent gene sets and edges represent the 'gene overlap' between gene sets measured by the overlap 'score'. By default, the overlap 'score' combined an equal-weighted Jaccard coefficient (50%) and overlap coefficient (50%), and the cutoff the overlap 'score' is 0.375. Deep learning network provides layouts using extracted features or embeddings from networks (Grover and Leskovec, 2016; Muzio *et al.*, 2021; Perozzi *et al.*, 2014; Tang *et al.*, 2015). However, none of the current layout algorithms integrate different types of biological factors into the layout generation and synthetically present a global view of genomic information.

Based on the above considerations, we propose a new layout algorithm, the distance-bounded energy-field minimization algorithm (DEMA). DEMA aims to incorporate biological network properties and prior knowledge information of the network to enhance network biology explorations and applications. It generates a layout where network hubs and the functional modules can be shown visually with controlled parameters including hub shrinking (HS) and the space filling (SF). A DEMA-optimized network layout may help users identify enriched biological pathways and enriched gene sets in a network context. In addition, the DEMA network layout may reveal gene signature patterns derived from functional genomics analysis using network visualization tools such as GeneTerrain (You *et al.*, 2010). The essential input of DEMA is a PPI network that describes the physical or co-expression relationships between genes. The biological factors analyzed by computational measures are categorized and defined in our layout. Interaction coefficient (IC) is defined as a measure to evaluate the reliability of the edges in the PPI network using statistical methods (Chen *et al.*, 2017). The IC can be extended to any biological feature that reflects the edge strength. Effect coefficient (EC) is defined as a measure to evaluate the correlation between genes based on gene expressions, such as the Pearson correlation coefficient (Langfelder and Horvath, 2008). EC can be substituted for any metric representing any correlation between the nodes in networks. Fold change (FC) is defined as the fold change of the gene expressions from case samples to control samples (Smyth, 2005). FC can be extended to any read-out values from experiments. DEMA can take one or more of IC, EC and FC as inputs and produce a layout by using the length of the edge to present these biological factors. Furthermore, the layout also considers gene-set annotation. By denoting which genes share similar functions, DEMA can group these genes together, and it can be extended to any grouping annotation of nodes.

The layout generated by DEMA has four properties:

1. *Topological property*: The layout generated by DEMA shows some modularities so that the hubs in the network can be observed. The layout generated by our method should make sure that the nodes do not overlap each other and can be seen clearly.
2. *Edge-weighted property*: If the edge that connects two nodes has a strong relation (EC or IC), then these two nodes will become close; otherwise, they are far away from each other.
3. *Node-weighted property*: Two nodes that share an edge will become close if they both have high node weights (FC) and have a positive association; otherwise, they are far away from each other.
4. *Grouping ability*: Gene-set annotation analysis is a popular technology to functionally analyze the gene sets in the network. Gene sets are defined as clusters of genes organized by different functionalities. The layout generated by DEMA can present the gene-set relation by grouping the genes that share similar functions.

Table 1 shows the comparison between DEMA and popular layouts provided by Cytoscape (Shannon *et al.*, 2003) and Gephi (Bastian *et al.*, 2009) software. DEMA is the only one layout algorithm that includes four kinds of properties.

2 Materials and methods

2.1 Model definition

Energy function-based layout methods are the commonly used methods for designing a layout. They model a graph layout as a physical system where nodes are attracted and repelled by different kinds of forces. We introduce a parameterized energy model where nodes are repelled by the network topology and attracted by a few biological factors, i.e., interaction coefficient (IC), effect coefficient (EC) and fold change (FC) of gene expression. The model is also suitable for a binary network where FD algorithms are usually applied. In the following sections, we discuss the base model, parameterized energy model and parameterized energy model with grouping. The base model corresponds to the DEMA layout with basic topological property (L_0). Parameterized energy model corresponds to the DEMA layout with additional edge-weighted property (L_e) and additional node-weighted property (L_n). Parameterized energy model with grouping corresponds to the DEMA layout with additional gene-set grouping ability (L_s).

2.2 Parameterized energy model

The energy function is composed of repulsion energy and attraction energy, which is defined as follows:

$$\begin{aligned} E &= K_a E_a + K_b E_b \\ &= \sum_i \sum_j \frac{K_a * R_i * R_j}{d_{i,j}} \\ &\quad + \sum_{i,j \in NET} \left(\frac{K_b * FC_i * FC_j}{w_{i,j} - d_{i,j}} - \frac{K_b * FC_i * FC_j}{w_{i,j} - D_{i,j}} \right), \end{aligned} \tag{1}$$

where $0 < D_{i,j} \leq d_{i,j} < w_{i,j}$ and $D_{i,j}$ is an extremely small constant (10^{-8} in default) that makes sure any two nodes will not collapse. E_a denotes the repulsion energy and E_b denotes the attraction energy. In the above equation, NET denotes the PPI network, K_a, K_b are control coefficients. When K_b/K_a is large, the attraction influence is strong to pull the nodes together, while the nodes are pushed away when K_b/K_a is small. R_i is the RP score of gene i, FC_i is the fold change (FC) of gene i, $d_{i,\ j}$ is the distance between gene i and gene j, and $w_{i,\ j}$ is a constraint weight to make the distance $d_{i,\ j}$ is smaller than $w_{i,\ j}$.

Specifically, RP score R_i proposed in Chen *et al.* (2006) separates the important nodes. It is defined as follows:

$$R_i = e^{k * ln\left(\sum_{j, where(i,j) \in NET} IC_{i,j}\right) - ln \sum_{j, where(i,j) \in NET} 1}, \tag{2}$$

where $IC_{i,j}$ is the interaction coefficient (IC) between gene i and gene j. k is set to two by default. Here, the PPI score between gene i and gene j is used as an interaction coefficient. FC_i is the fold change (FC) of gene i expression values between the control samples and the case samples.

Constraint weight $w_{i,\ j}$ is defined as

$$\mathrm{w}_{i,j} = \max\left(1 / \left(-\log\left(\frac{1 - IC_{i,j}}{1 + IC_{i,j}}\right) - \log\left(\frac{1 - EC_{i,j}}{1 + EC_{i,j}}\right) + K_{i,j}\right),\ D_{i,j} + \varepsilon\right), \tag{3}$$

where $K_{i,\ j}$ is set to the default of 1 and $EC_{i,\ j}$ is the effect coefficient (EC). Here, we use as EC the Pearson correlation of the genes. As the ranges of IC and EC are both [−1,1], the range of parameter w is (0, 1] due to the constraint of $D_{i,j} + \varepsilon$.

2.3 Base model

In some situations, we cannot get the FC, IC and EC. When this information is missing, the default values are set. For the attraction energy E_b, FC, IC and EC are all set to be 1. Because $1/\left(-\log\left(\frac{1 - IC_{i,j}}{1 + IC_{i,j}}\right) - \log\left(\frac{1 - EC_{i,j}}{1 + EC_{i,j}}\right) + K_{i,j}\right)$ tends to 0 when the IC or EC tends to 1, $w_{i,j} = D_{i,j} + \varepsilon$.

In the repulsion energy E_a, due to IC equaling 1, the RP score is defined as

Table 1. The comparison of different layout algorithms

Layout	Topological property	Node-weighted property	Edge-weighted property	Gene-set grouping property
OL (Cytoscape)	✓	✗	✗	✗
FA2 (Gephi)	✓	✗	✗	✗
FR (Gephi)	✓	✗	✗	✗
FD (Cytoscape)	✓	✗	✗	✗
EWFR (Cytoscape)	✗	✗	✓	✗
EMap (Cytoscape)	✗	✗	✗	✓
DEMA	✓	✓	✓	✓

$$R_i = e^{k*\ln\left(\sum_{j,where(i,j)\in NET} 1\right) - ln \sum_{j,where(i,j)\in NET} 1}. \tag{4}$$

It approximates the degree of the node. In the base model, the energy function is written as

$$E = \sum_i \sum_j \frac{K_a * R_i * R_j}{d_{i,j}} + \sum_{i,j\in NET} \left(\frac{K_b}{D_{i,j} + \epsilon - d_{i,j}} - \frac{K_b}{\epsilon} \right). \tag{5}$$

2.4 Parameterized energy model with grouping

Because gene-set analysis grows in popularity and increases the interpretability of data, we further improve our energy model by combining it with the gene set.

The gene-set energy is added into our model as the third item, which is defined as follows:

$$E_c = \sum_{s\in \text{gene sets}} \sum_{i\in s} d_{i,c}^2, \tag{6}$$

where s is a gene set, and c is the center position of s. Therefore, the energy function is

$$E = K_a E_a + K_b E_b + K_c E_c. \tag{7}$$

2.5 Complete model of DEMA

The complete model of DEMA is

$$\begin{aligned} E &= K_a E_a + K_b E_b + K_c E_c \\ &= \sum_i \sum_j \frac{K_a * R_i * R_j}{d_{i,j}} \\ &+ \sum_{i,j\in NET} \left(\frac{K_b * FC_i * FC_j}{w_{i,j} - d_{i,j}} - \frac{K_b * FC_i * FC_j}{w_{i,j} - D_{i,j}} \right) \\ &+ K_c \sum_{s\in \text{gene sets}} \sum_{i\in s} d_{i,c}^2. \end{aligned} \tag{8}$$

2.6 Relation between parameter *w* and IC as well as EC

When IC or EC increases, the parameter w decreases (Supplementary Fig. S1), such that it can decrease the maximum distance between them. It is because of the constraint that the distance d between two genes should be smaller than their corresponding w. If a pair of genes have high IC and EC, which denote a strong relation, the pair should be close to each other.

2.7 Relation between the small energy system and parameter *w*

Assume that there are just two genes forming a small energy system composed of the repulsion energy and the attraction energy. If the other parameters in the energy system are fixed except for parameter w, when the distance is increasing, the repulsion energy decreases while the attraction energy increases. This results in an optimal distance between genes such that the total energy is minimized. Meanwhile, w will influence the optimal distance. When w is small, the optimal distance shrinks (Supplementary Fig. S2).

2.8 Relation between the group energy and distances from nodes to their center

For each group, when the distance between the node and the center increases, the group energy will increase. Minimizing the group energy pulls the nodes in the group close together and attracts each node close to the center of the group (Supplementary Fig. S3).

2.9 Finding a greedy solution

2.9.1 Solution for parameterized energy model

Given the coordinates of a gene i in the 2D Euclidean plane by $p_i = (x_i, y_i)$, the distance $d_{i,\ j}$ between gene i and j is written as follows

$$d_{i,j} = \sqrt{(p_i - p_j)^2} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \tag{9}$$

The goal of the algorithm is to find values for the coordinates of each node i to minimize the energy function $E(p_1, p_2, \ldots, p_m)$. It is a constrained optimization problem; however, because the force is the derivative of potential energy (Kobourov, 2012), we can calculate the force exerted on one node and move it at a time to solve this problem. With respect to the position of one node i, we compute the gradient of the energy function and calculate the force exerted by other nodes.

For each gene i, the strength of the repulsion force exerted by node j is

$$f_{i,j}^a = \frac{K_a * R_i * R_j}{d_{i,j}^2}. \tag{10}$$

And the direction is -(p_j-p_i). While the strength of the attraction force exerted by node j is

$$f_{i,j}^b = \frac{K_b * FC_i * FC_j}{(w_{i,j} - d_{i,j})^2}. \tag{11}$$

And the direction is p_j-p_i. Therefore, the total force is

$$f = -\sum_i \sum_j f_{i,j}^a (p_j - p_i) + \sum_{(i,j)\in NET} f_{i,j}^b (p_j - p_i). \tag{12}$$

After we get the direction of the node i moved by the force, the step length is to reduce the total energy. The step length is chosen to be 0.001 from a set L ranged from 0.001 to 1 in the greedy algorithm, which means to choose the maximum element $l \in L$ such that the constraint will not be violated and the total energy will decrease after moving the point i with such a step length. The greedy algorithm can be implemented by binary search, resulting in a fast algorithm to find the result with O(log |L|).

2.9.2 Solution for parameterized energy model with grouping

To solve the above model, we use the *EM* (*expectation and maximization*) algorithm to find the local optimal solution iteratively.

In the expectation stage, we fix the positions of the genes and determine the position of the center of the gene set according to the mean value of the positions of the corresponding genes, which makes the energy of the gene-set minimal.

In the maximization stage, we also move each gene according to the force exerted on it. For each gene i, the strength of the gene-set force exerted by its gene set(s) is

$$f_i^c = K_c \sum_{s \in S} 2d_{i,c}, \tag{13}$$

where S is a set composed of the gene set containing gene i and c is the center of the gene set.

And the direction is $p_c - p_i$. Therefore, the total force is

$$f = -\sum_i \sum_j f_{i,j}^a (p_j - p_i) + \sum_{(i,j) \in NET} f_{i,j}^b (p_j - p_i) + \sum_i f_i^c (p_c - p_i) \tag{14}$$

After we get the direction of the point i to move on by the force, the step length is set to reduce the total energy. The step length is chosen to be 0.001 from a set L range from 0.001 to 1 in the greedy algorithm.

2.10 Heuristic start

To further converge the energy function to a minimum fast, a heuristic start is proposed. First, we construct the initial network by connecting each pair of nodes that share an edge.

Two kinds of energy compose the small energy system of each pair of nodes that share an edge in the formula (1).

From the derivation, we can find the optimal distance to minimize the energy system is

$$d = \frac{w\sqrt{K_a * R_i * R_j}}{\sqrt{K_b * FC_i * FC_j + K_a * R_i * R_j}} \tag{15}$$

Then, we use the shortest path to connect all of the nodes to build a complete network.

After we define the distances between any pair of nodes, the formulation is transformed into the Kamada–Kawai (KK) model (Kamada and Kawai, 1989; Khoury *et al.*, 2012), and we can solve it according to the stress majorization algorithms (Gansner *et al.*, 2004; Khoury *et al.*, 2012). It is proved that it needs a few iterations to achieve an appealing layout by using the stress majorization. However, the layout cannot guarantee that the empirical distance between the nodes which share an edge is smaller than the edge weight since the empirical distance is sometimes larger than the ideal distance. To make the initial layout satisfy this constraint, we constantly replace K_B by 2 * K_B so that the optimal distance becomes smaller until the empirical distance between the nodes which share an edge is smaller than the edge weight. The layout generated by the heuristic start (HS) can be used as an initial layout for DEMA. Compared with random start (RS), DEMA with HS can show a layout with fewer iterations (Supplementary Fig. S4).

2.11 Network generation

The random networks are generated by NetworkX package (available from https://networkx.github.io/) in Python. They satisfy the scale-free property. The distribution of the degrees of the 100 nodes, 500 nodes and 1000 nodes are shown in Supplementary Figure S5, respectively.

3 Results

DEMA is a layout algorithm implemented as an easily operated Cytoscape plugin (download at http://discovery.informatics.uab.edu/dema). The input to DEMA consists of two files: the edge file that describes the attributes of the edges connecting the nodes and the node file that describes the properties of the nodes. As depicted in Figure 1, there are four properties of DEMA corresponding to four parts (A, B, C, D) of inputs, respectively. Part A is a PPI network and the necessary input to DEMA. Other parts are optional. The four properties will be introduced one by one. A case study is given at the end.

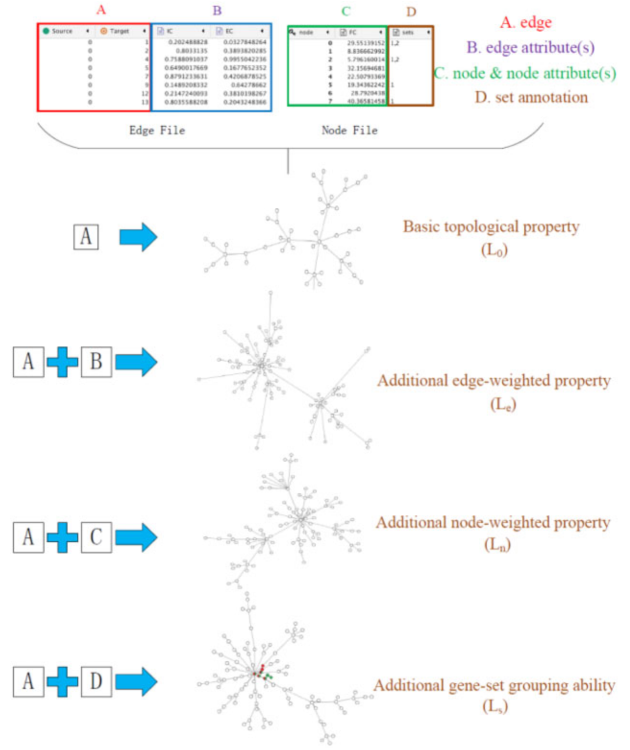

A
B
C
D
A. edge
B. edge attribute(s)
C. node & node attribute(s)
D. set annotation
Edge File
Node File
A
Basic topological property
(L_0)
A
B
Additional edge-weighted property
(L_e)
A
C
Additional node-weighted property
(L_n)
A
D
Additional gene-set grouping ability
(L_s)

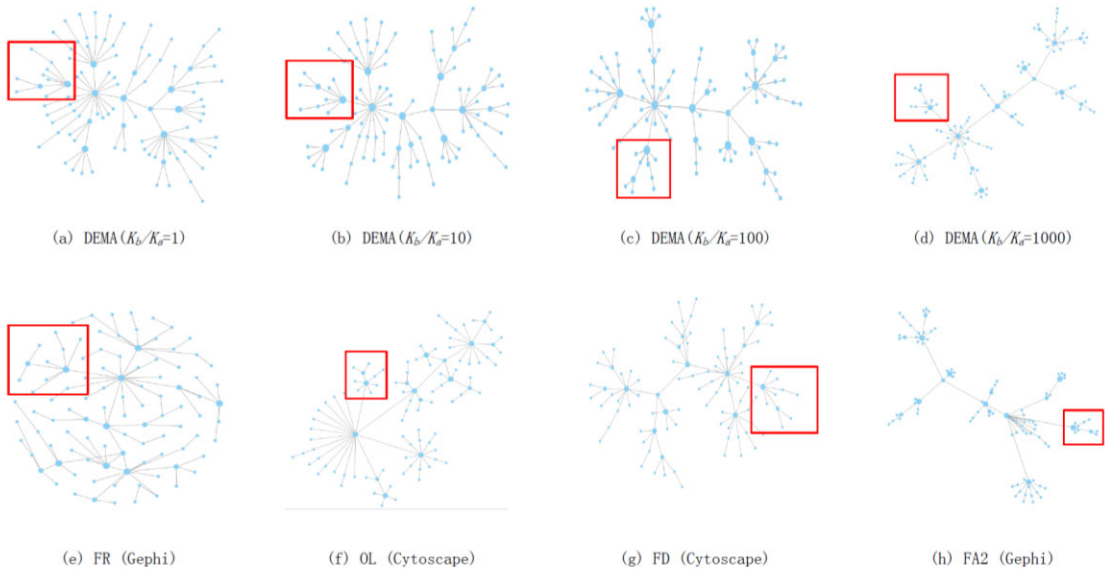
Fig. 1. Illustration of DEMA. There are two files (node file and edge file) and four parts of them. Part A describes the PPI network. Part B describes the attributes of the edges. Part C describes the attributes of the nodes. Part D annotates the sets that the nodes belong to. Among them, Part A is the essential input and combines with other parts to perform additional corresponding functional properties

3.1 Topological property

As depicted in Figure 1, by importing the PPI network that uses edges to describe the network, DEMA can generate a layout with topological modularity (L_0).

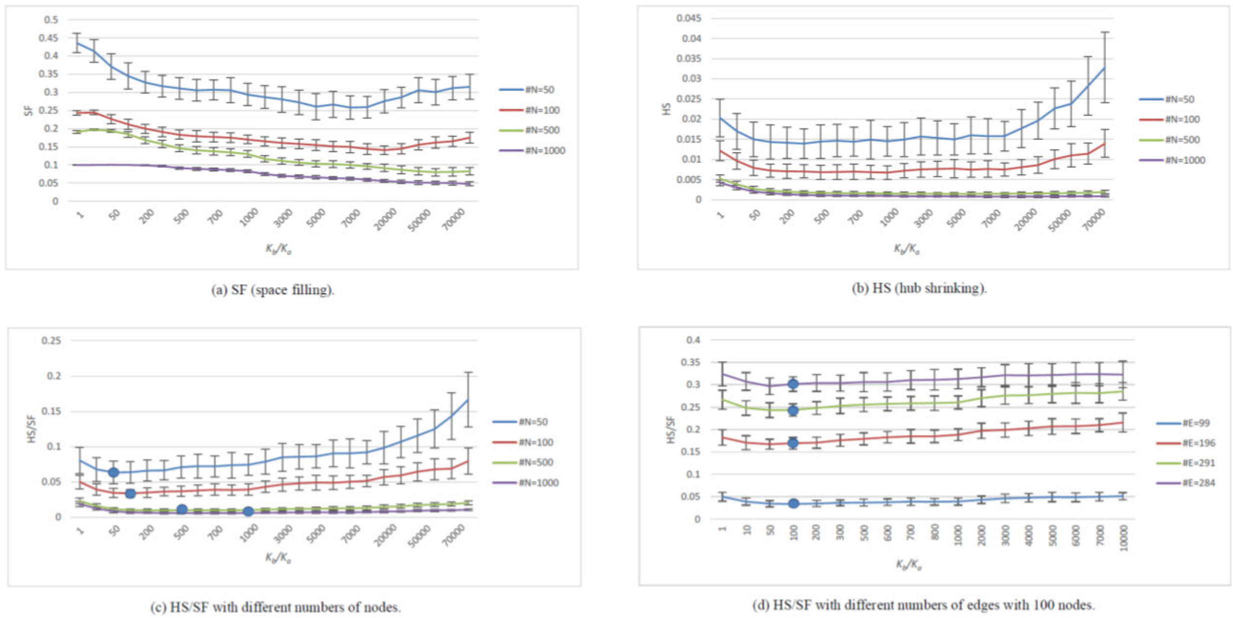
The comparison between DEMA and other layouts with a random network of 100 nodes satisfying scale-free property (Barabási, 2009) is shown in Figure 2. Cytoscape and Gephi provided the best-performed layout with their default parameters that we adopted in the comparison. Figure 2(a–d) is generated by DEMA with different values of control the parameter K_b/K_a. Figure 2(e–h) are the popular layouts provided by software Cytoscape and Gephi. They are FR model (Fruchterman and Reingold, 1991), FD layout (http://www.prefuse.org), Organic layout (OL) (Cline *et al.*, 2007) and ForceAtlas2 (FA2) (Jacomy *et al.*, 2014), respectively. From Figure 2(a–d), we can see that when K_b/K_a becomes large, the modularity in the layout is more obvious. However, the distance between the nodes surrounding the hubs and the hubs becomes small, which makes it difficult to observe the surrounding nodes. Compared with other layouts, we can see that (e) subfigure roughly corresponds to (a), and the hubs in the layouts are both not obvious. (f) roughly corresponds to (b), while there is no hierarchy in the hub and its neighborhood in (f). In Figure 2b, we can see that the edge between the small hub and the big hub is longer than the edges between other nodes and the big hub in the crop regions in the red rectangle frames. In contrast, in Figure 2f, the edge between the small hub and the big hub is the same length as the edges between other nodes and the big hub. (g) roughly corresponds to (c), while some edges cross in (g). (h) roughly corresponds to (d). Both of them clearly show the hubs, but the nodes cannot be observed clearly.

To choose a proper parameter K_b/K_a for DEMA, we access the network modularity by evaluating visual clarity and network modularity. With a proper parameter K_b/K_a, DEMA can clearly observe the hubs and the nodes. We use space filling (SF) to evaluate the visual clarity and hub shrinking (HS) to evaluate the structural modularity. First, we normalize the nodes' coordinates at [0, 1] for each


(a) DEMA (K_b/K_a=1) (b) DEMA (K_b/K_a=10) (c) DEMA (K_b/K_a=100) (d) DEMA (K_b/K_a=1000)

(e) FR (Gephi) (f) OL (Cytoscape) (g) FD (Cytoscape) (h) FA2 (Gephi)

Fig. 2. A comparison of different network layouts with network graph node size $n = 100$. (**a–d**) The layouts drawn by DEMA with different parameters K_b/K_a are described. (**e–h**) The layouts drawn by different layout algorithms



(a) SF (space filling).

(b) HS (hub shrinking).

(c) HS/SF with different numbers of nodes.

(d) HS/SF with different numbers of edges with 100 nodes.

Fig. 3. Effects of layout by varying network sizes and DEMA K_b/K_a parameters. #N denotes the number of the nodes. #E denotes the number of the edges. (**a**) Space-filling (SF) is used to evaluate visual clarity. The larger, the better. (**b**) Hub shrinking (HS) is used to evaluate structural modularity. The smaller, the better. (**c**) and (**d**) use HS/SF to achieve a balance between the visual clarity and the structural modularity. The smaller, the better. The round dots in (c) indicate the optimal K_b/K_a values with the lowest of HS/FS ratio. The round dots in (d) indicate the optimal choice for K_b/K_a determined by the input node size

axis. Then, we build a grid and see what percentage of the small squares in the grid are occupied by the nodes; it is the SF measure. The larger the percentage, the more dispersed the nodes are. For the HS measure, the coordinates of the nodes are also normalized at first. We then calculate the interaction arithmetic mean of square distance between each pair of nodes that share an edge. Obviously, the smaller the HS value, the more the nodes shrink to the hubs. The networks with 50 nodes, 100 nodes, 500 nodes and 1000 nodes are randomly generated. Each kind of network is generated 100 times. SF and HS are calculated and shown in Figure 3a and b. For SF, the number of the small squares in a grid is almost twice the number of the nodes. Although there are some unstable situations, HS and SF generally decrease as K_b/K_a increases. SF stabilizes after it reaches a valley, while HS bounces up after K_b/K_a crosses a threshold. The reason that HS bounces up is that the distance between the hubs increases, and it affects HS more than the distance between the hub and the peripheral nodes does. To achieve a balance between SF and HS, we calculate HS/SF. The result is shown in Figure 3c. From the figure, we can see that the best parameter for the smallest HS/SF is usually close to the number of nodes. Also, from Figure 3d, we can

Table 2. HS/SF for different layouts

	FR	FD	OL	FA2	DEMA (K_b/K_a = 100)
HS	0.026	0.010	0.015	0.008	0.008
SF	0.210	0.418	0.382	0.219	0.352
HS/SF	0.052	0.025	0.040	0.036	**0.023**

The bold value indicates the DEMA performed the best (the lowest value in the HS/SF metric).

observe that given 100 nodes with different numbers of edges, the best parameter for the smallest HS/SF is still close to the number of nodes. Therefore, the default value of K_b/K_a is the number of nodes.

The HS/SF values for the layouts shown in Figure 2 are calculated in Table 2. From the results, we can see that DEMA achieves the lowest value and produces the best balance between the hub shrinking (HS) and the space-filling (SF).

3.2 Edge-weighted property

In Figure 1, Part B of the edge file includes the two computational biological factors, interaction coefficient and effect coefficient, as the attributes of the edges. Interaction coefficient (IC) is defined as a measure to evaluate the reliability of the edges in the PPI network using statistical methods [11]. Effect coefficient (EC) is defined as a measure to evaluate the correlation between genes based on gene expressions, such as the Pearson correlation coefficient [12]. By importing the PPI network and these biological factors represented as the edge weights, the layout in the figure not only shows the network modularity but also reflects the relationship between the nodes by the length of the edge (L_e). If any two nodes that share an edge have a strong IC or EC, they are close to each other; otherwise, they are far away from each other.

We use the P-value to show that the length of the edge can demonstrate the degree of the relationship. For IC and EC, the edges are sorted according to their values from largest to smallest. We select as top edges the top p percent of the total edges and calculate the average length of the top edges. Then, we randomly select p percent of the edges and check whether the average length of the selected edges is smaller than the average length of the top edges. The selection is repeated 10 000 times. The P-value is calculated as the probability that the average length of the randomly selected edges is smaller than the average length of the top edges. We generate six kinds of random networks, and the IC and EC values of the networks are randomly generated 100 times. These six kinds of random networks are network 1 (100 nodes and 99 edges), network 2 (100 nodes and 196 edges), network 3 (100 nodes and 291 edges), network 4 (500 nodes and 499 edges), network 5 (500 nodes and 996 edges) and network 6 (500 nodes and 1491 edges). From Supplementary Figures S6 and S7, we see that the P-values of EC and IC are below 0.05 for the network with 100 nodes and 500 nodes. They are significant. It means that our parameterized layout not only has network modularity but also reflects the relationship between any two nodes that share an edge by the length of the edge.

3.3 Node-weighted property

In Figure 1, Part C of the node file includes fold change, a biological factor, as the attribute of the node. Fold change (FC) is defined as the fold change of the gene expressions from case samples to control samples (Smyth, 2005). By importing the PPI network and the FC of nodes, the layout in the figure not only shows the network modularity but also reflects the relation between the nodes by the length of the edge (L_n). If any two nodes that share an edge both have the high FC, they are close to each other.

We also use the P-value to show that the length of the edge can demonstrate the degree of the relationship. If two nodes that share an edge both have a high FC value, they should be close to each other; otherwise, they are far away from each other. The value of the edge is defined as the product of the values of the two nodes that the edge connects. The edges are then sorted from largest to smallest. We select as top edges the top p percent of the total edges and calculate the average length of the top edges. Then, we randomly select the p percent of the edges and check whether the average length of the selected edges is smaller than the average length of the top edges. The selection is repeated 10 000 times. The P-value is calculated as the probability that the average length of the randomly selected edges is smaller than the average length of the top edges. We generate a random network, and the values of FC for the networks are randomly generated 100 times. From Supplementary Figure S8, we see that the P-values of FC are all below 0.08 no matter for the network with 100 nodes or 500 nodes. They are significant.

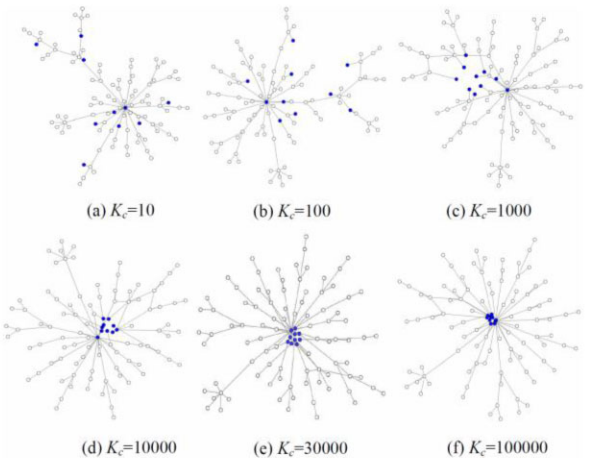

(a) K_c=10 (b) K_c=100 (c) K_c=1000
(d) K_c=10000 (e) K_c=30000 (f) K_c=100000

Fig. 4. Effects of DEMA layout by varying *Kc*. The solid nodes represent the nodes in the set

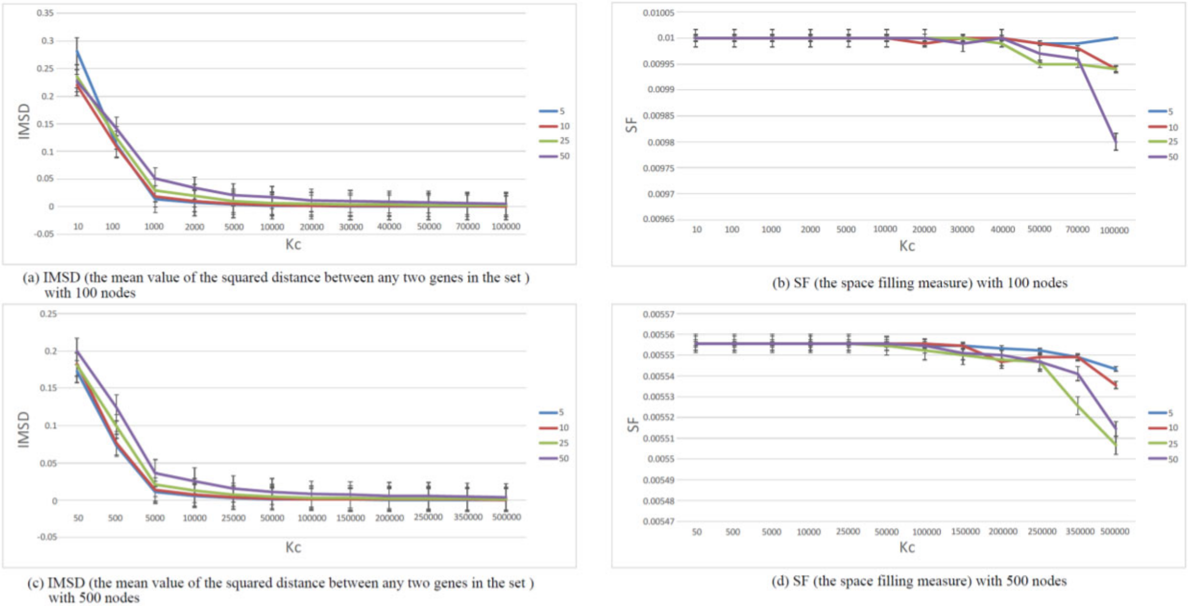
3.4 Gene-set grouping ability

In Figure 1, Part D of the node file includes gene-set annotation, denoting what gene sets each gene is from. By importing the PPI network and the gene-set annotation, the layout can show the additional grouping ability (L_s). From the figure, one can see that in some rows of the set column, multiple sets are denoted and separated by commas. Some genes may belong to multiple sets. DEMA will group the genes in the same set. In DEMA, there is a parameter K_c to control the degree to which the genes in the same set are grouped. As described in Figure 4, the genes in the set will become closer as K_c increases.

To evaluate the grouping performance, we define two measures, the inside mean value of the squared distance (IMSD) and space filling (SF). IMSD is defined as the mean value of the squared distance between any two genes in the set divided by the mean value of the pairwise squared distance among all the input genes. SF is defined in the above part, which is used to evaluate visual clarity. We generate 10 random networks with 100 nodes and calculate the IMSD and SF for four gene sets. They include gene subsets (5, 10, 25 and 50%) randomly selected from all the genes, respectively. The averaged results are shown in Figure 5a and b. From Figure 5a, we can see that IMSD becomes smaller as K_c increases, which shows that the nodes in the group become closer. From Figure 5b, we can see that SF is almost the same before K_c is equal to 40 000. When K_c is larger than 40 000, SF declines rapidly, which means that some nodes overlap. We generate 10 random networks with 500 nodes and calculate the IMSD and SF for four gene sets. The averaged results are shown in Figure 5c and d. In Figure 5c, IMSD also becomes smaller as K_c increases. And in Figure 5d, SF is almost the same before K_c is equal to 150000. When K_c is larger than 150 000, SF also declines rapidly. Therefore, the default K_c value is defined as 300 times as the number of nodes to make a balance between the visual clarity and the grouping degree.

3.5 Case study of autism spectrum disorder

Autism spectrum disorder (ASD) is analyzed by DEMA as a case study. Autism is a complex disease, and it is still difficult to



(a) IMSD (the mean value of the squared distance between any two genes in the set) with 100 nodes

(b) SF (the space filling measure) with 100 nodes

(c) IMSD (the mean value of the squared distance between any two genes in the set) with 500 nodes

(d) SF (the space filling measure) with 500 nodes

Fig. 5. DEMA gene-set grouping ability with varying K_c. In (**a**) and (**c**), (IMSD) is defined to measure the distance between the nodes in the set. In (**b**) and (**d**), SF is defined to measure the space filling for all the nodes. The legend represents the gene subsets (5, 10, 25 and 50%) randomly selected from all the genes, respectively

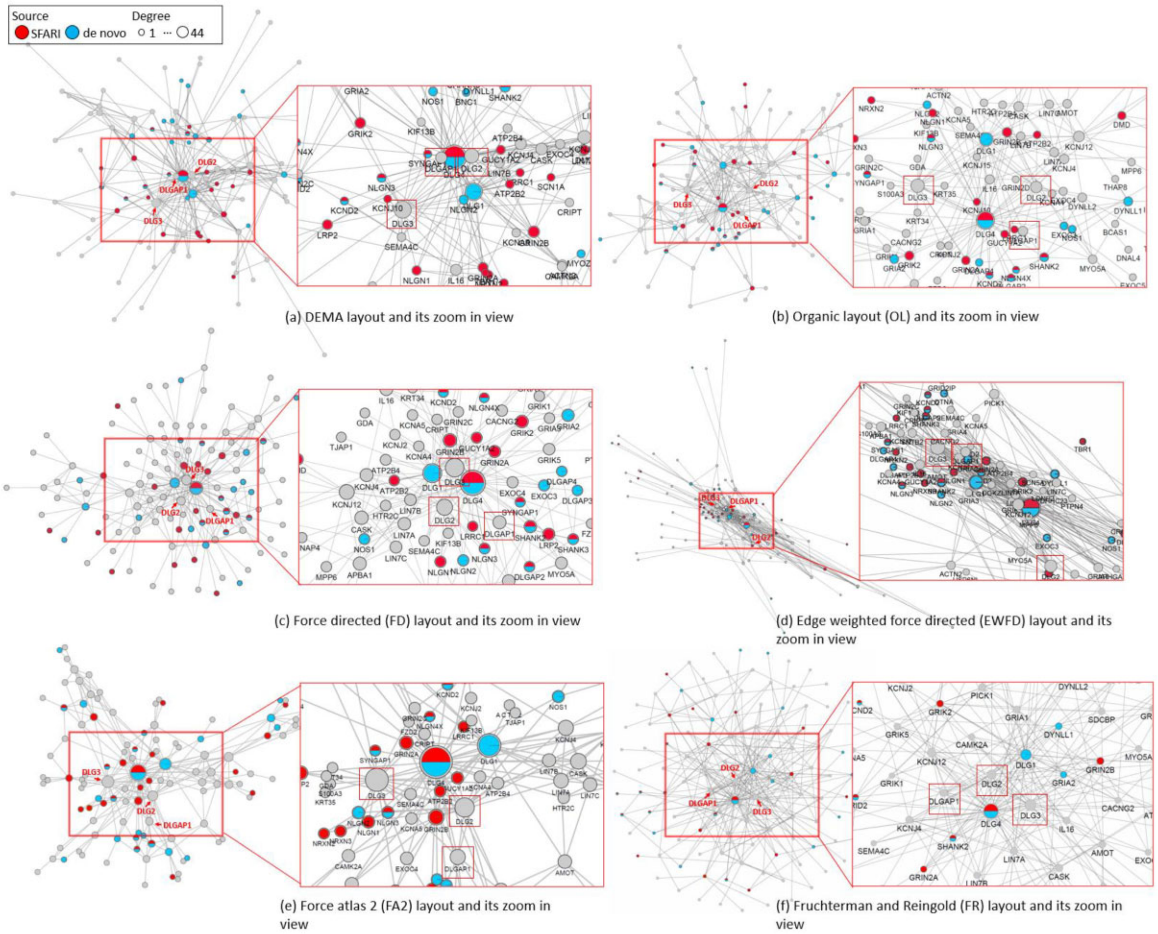
determine its etiology. In the paper Li *et al.* (2014), a human PPI network is built and decomposed into 817 topological modules based on the human protein interactome from BioGrid (Stark *et al.*, 2011) (Supplementary File S1). Among these modules, Module #13 shows significant enrichment related to ASD and includes 199 genes. In this module, the SFARI reference ASD genes and *de novo* CNVs are denoted (Supplementary File S1). We took the cluster #13's genes and extracted the subset of the PPIs associated to those genes in the cluster #13 directly from the supplemental of the paper (Stark *et al.*, 2011). This module is used to draw a layout by DEMA, organic layout (OL), FD, edge-weighted force-directed (EWFD), Forcealtas2 (FA2) and Fruchterman and Reingold (FR) layout (Fig. 6), respectively. The nodes in red are SFARI reference genes, and the nodes in blue are *de novo* CNVs. The other layouts only show which genes are connected to the related genes. However, the layout by DEMA can group the related genes together and observe the hubs that connect to these related genes. DEMA performances the best in both SFARI and *de novo* gene groupings measured by IMSD scores (Supplementary Table S1). Three candidate genes are denoted by the arrows, which will be further analyzed in the following. One can see that DEMA can group the genes in the same set together. By zooming in DEMA, in Figure 6a, DLG2 (Jiangxie *et al.*, 2014), DLG3 (Kantojärvi *et al.*, 2011) and DLGAP1 (Egger *et al.*, 2014) are highly connected to the ASD-related genes, which are also related to ASD according to the literature. These genes are considered to be new candidate genes related to ASD.

3.6 Case study of Alzheimer's disease

To illustrate the biologically functional gene groups in a heterogeneous disease like Alzheimer's disease (AD), an important parameter K_c in DEMA has been designed to group and visualize the acknowledged-based gene sets in disease gene networks. In our study, 680 candidate genes in view of genetic risk were collected from the AlzGene database (https://www.alzforum.org/alzgene). To construct the gene network, the PPI were retrieved from the HAPPI-2 database (Chen *et al.*, 2017) using the quality more than or equal to 4-star (PPI score $\geq$ 0.75), and the PPI scores were used as IC scores (Supplementary File S2). There were 558 genes connected by 3459 PPIs. With the aim of highlighting the important genes in the network. After proceeding with the gene enrichment analysis using PAGER 2.0 database (Yue *et al.*, 2015, 2018), we found 6 highly relevant pathways using the false discovery rate (FDR) $\leq$ 0.05, overlaps $\geq$ 15 and similarity score $\geq$ 0.1. The similarity scores were calculated using the methods in (Huang *et al.*, 2012) (Table 3). After setting the parameters $K_b/K_a = 558$ and $K_c = 150\,000$ suggested by the DEMA, the AD gene network was visualized as well as the crucial genes in pathways were grouped with clear functional patterns in Figure 7a compared to the organic layout in Figure 7b. Several key genes with high-genetic risk were revealed by overlapping four or five pathways in Figure 7, Supplementary Table S2 and File S2. For instance, APOA1, APOB, LDLR, APOE and ABCA1 involve in at least four pathways, 'Vitamin B12 metabolism', 'Folate metabolism', 'Plasma lipoprotein assembly, remodeling, and clearance', 'Statin inhibition of cholesterol production' and 'Retinoid metabolism and transport'. In DEMA, these genes were grouped coherently. TNF, IL6, IL1B, CCL2 and ICAM1 are located near each other, and these genes are involved in the three pathways, 'Vitamin B12 metabolism', 'Folate metabolism' and 'Interleukin-10 signaling'. ALB locates near APOE, and it shares three identical pathways with APOE. APOA2, APOC3, APOA4 and APOC2 locate near to each other. These genes share the three pathways 'Plasma lipoprotein assembly, remodeling, and clearance', 'Statin inhibition of cholesterol production' and 'Retinoid metabolism and transport'. SCARB1 locates at the intersection of the pathway group, 'Vitamin B12 metabolism', 'Folate metabolism' and 'Plasma lipoprotein assembly, remodeling, and clearance'. In all, DEMA has successfully highlighted and grouped the functional genes together with gene-set-based information, which possessed considerable application potential for further investigation.

4 Conclusion and discussion

DEMA introduced a parameterized energy model to integrate the critical biological factors, the biological network analysis, the strength of PPIs, the gene-to-gene correlations, the gene strength and the functional gene groups. To find a local optimal solution iteratively, we applied EM (expectation and maximization) algorithm. Thus, DEMA delivers sub-optimal solutions to the global problem, and the local-field energy needs to be minimized using the stacked forces to continue the search beyond local optimality. To search for the shortest paths in the heuristic start, we applied the stress majorization algorithms, and it is faster than the random start to get the



(a) DEMA layout and its zoom in view

(b) Organic layout (OL) and its zoom in view

(c) Force directed (FD) layout and its zoom in view

(d) Edge weighted force directed (EWFD) layout and its zoom in view

(e) Force atlas 2 (FA2) layout and its zoom in view

(f) Fruchterman and Reingold (FR) layout and its zoom in view

Fig. 6. Biological interpretability for ASD (autism spectrum disorders). The blue nodes indicate the genes from *de novo* CNVs and the red nodes indicate the gene from SFARI reference. DLG2, DLG3 and DLGAP1 are highly connected to the ASD-related genes. (**a**) Organic layout, (**b**) DEMA layout, (**c**) Force-directed (FD) layout, (**d**) Edge-weighted force-directed (EWFD) layout, (**e**) Force atlas 2 (FA2) layout and (**f**) Frutchterman and Reingold (FR) layout (A color version of this figure appears in the online version of this article)

Table 3. The information of enriched PAGs from wiki-pathway in Alzheimer's disease

ID	Name	Size	Overlap	Similarity	Adj.P	Color
WAG002004	Vitamin B12 metabolism (WP1533)	53	29	0.142	3.84E−38	Red
WAG002021	Folate metabolism (WP176)	72	30	0.116	7.30E−35	Orange
WAG002786	Plasma lipoprotein assembly, remodeling and clearance (WP4129)	67	25	0.102	2.64E−28	Yellow
WAG002857	Statin inhibition of cholesterol production (WP430)	31	19	0.144	1.92E−26	Green
WAG002729	Interleukin-10 signaling (WP4063)	46	20	0.109	3.08E−24	Blue
WAG002606	Retinoid metabolism and transport	32	15	0.11	7.05E−19	Purple

Note: The 'ID' represents the PAG ID, in which the details of PAG can be retrieved using the url http://discovery.informatics.uab.edu/PAGER/index.php/geneset/view/[PAG_ID]. The 'Overlap' represents the number of overlapped genes between the queried gene list and the PAG gene members. The 'similarity' is calculated based on the combination of overlap coefficient and Jaccard index. The 'Adj.P' represents the adjusted *P*-values.

initial coordinates. Additionally, to evaluate the structure modality in comparing DEMA to the other layout algorithms, we introduced space filling (SF) and hub shrinking (HS). DEMA takes the lead in balancing between the HS and SF. We statically validated the node, edge, and grouping properties in the synthetic networks. DEMA can be extended the adaptivity in substituting the designed biological factors (IC, EC, FC and geneset grouping) to any equivalent feature in other Omics, such as metabolomics network, drug–drug interaction network and hybrid networks across different biological domains. Further, we performed two real-world case studies using biological networks. Since DEMA adapts the energy-based algorithm, the basic model using an unweighted network may produce a layout similar to the organic layout. However, given the biologically functional annotations, DEMA makes a big difference by revealing biological groupings and those surrounding candidates intuitively. We expect DEMA to be a major tool in network analytics.

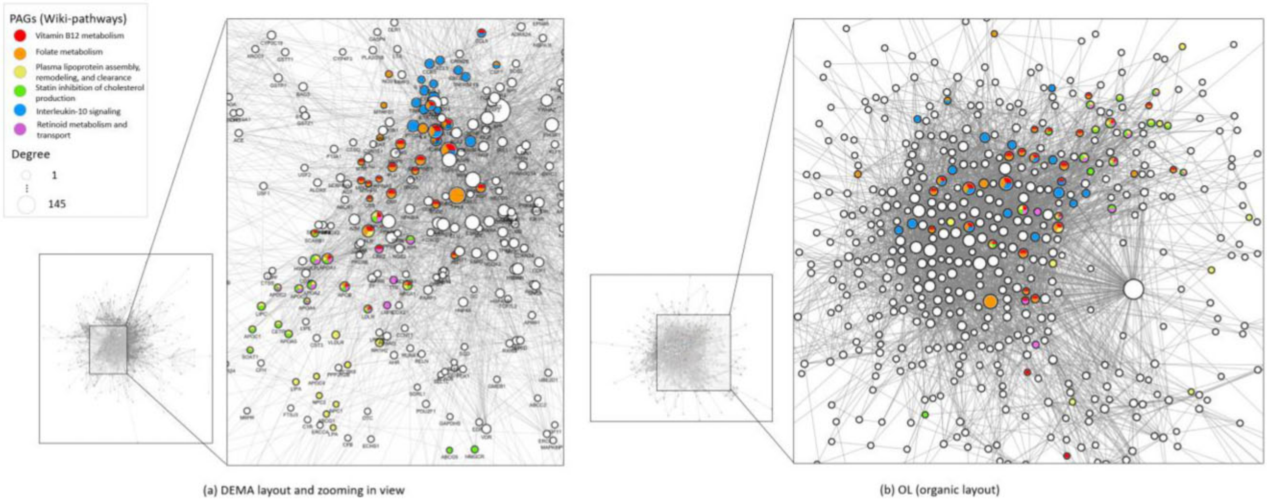

PAGs (Wiki-pathways)
● Vitamin B12 metabolism
● Folate metabolism
● Plasma lipoprotein assembly, remodeling, and clearance
● Statin inhibition of cholesterol production
● Interleukin-10 signaling
● Retinoid metabolism and transport
Degree
1
145
(a) DEMA layout and zooming in view
(b) OL (organic layout)

Fig. 7. Biological interpretability of Alzheimer's disease (AD). (**a**) DEMA layout with enriched PAG (wiki-pathway) annotations and zooming in view of DEMA layout. (**b**) Organic layout. The highlighted genes in different colors represent the members of PAGs

Acknowledgements

All authors thank the following general technical support that made case studies included for this work: Samuel Bharti and Saghapour Ehsan for the manuscript revision and suggestions. Jelai Wang for managing the data management and data analysis computing framework. Hiren Desai from UAB Information Technology groups for supporting the backend Oracle database 19c management. Also, a sincere thank International Teochew Doctors Association for the general support.

Funding

The work was in part supported by the internal University of Alabama at Birmingham research grants to Dr Jake Y Chen and the National Institutes of Health grant awards [U54TR001005] in which Dr Chen serves as a co-investigator.

Conflict of Interest: The authors claim no conflicts of interest.

Data availability

The data underlying this article are available in the article and in its online supplementary material. The software and demo can be freely accessed at http://discovery.informatics.uab.edu/dema. The source code of DEMA is available here https://github.com/aimed-uab/DEMA.

References

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

Barabási,A.-L. (2009) Scale-free networks: a decade and beyond. *Science*, **325**, 412–413.

Bastian,M. *et al.* (2009) Gephi: an open source software for exploring and manipulating networks, Vol. 8. In: *ICWSM, San Jose, California, USA*, pp. 361–362.

Bock,J.O. and Ortea,I. (2020) Re-analysis of SARS-CoV-2-infected host cell proteomics time-course data by impact pathway analysis and network analysis: a potential link with inflammatory response. *Aging (Albany NY)*, **12**, 11277–11286.

Chen,J.Y. *et al.* (2017) HAPPI-2: a comprehensive and high-quality map of human annotated and predicted protein interactions. *BMC Genomics*, **18**, 182.

Chen,J.Y. *et al.* (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. In: *Pacific Symposium on Biocomputing, Grand Wailea, Maui, Hawaii, USA*, pp. 367–378.

Chen,J.Y. *et al.* (2007) A systems biology approach to the study of cisplatin drug resistance in ovarian cancers. *J. Bioinform. Comput. Biol.*, **5**, 383–405.

Cline,M.S. *et al.* (2007) Integration of biological networks and gene expression data using cytoscape. *Nat. Protoc.*, **2**, 2366–2382.

Drier,Y. *et al.* (2013) Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci. USA*, **110**, 6388–6393.

Egger,G. *et al.* (2014) Identification of risk genes for autism spectrum disorder through copy number variation analysis in Austrian families. *Neurogenetics*, **15**, 117–127.

Enright,A.J. and Ouzounis,C.A. (2001) BioLayout – an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, **17**, 853–854.

Fruchterman,T.M. and Reingold,E.M. (1991) Graph drawing by force-directed placement. *Softw. Pract. Exp.*, **21**, 1129–1164.

Gansner,E.R. *et al.* (2004) Graph drawing by stress majorization. In: *International Symposium on Graph Drawing, New York City, New York, USA*. Springer, 239–250.

Grover,A. and Leskovec,J. (2016) node2vec: scalable feature learning for networks, Vol. 2016. In: *KDD, San Francisco, California, USA*, pp. 855–864.

Hamburg,M.A. and Collins,F.S. (2010) The path to personalized medicine. *N. Engl. J. Med.*, **363**, 301–304.

He,X. and Zhang,J. (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet.*, **2**, e88.

Huang,H. *et al.* (2012) PAGED: a pathway and gene-set enrichment database to enable molecular phenotype discoveries. *BMC Bioinformatics*, **13** Suppl 15, S2.

Jacomy,M. *et al.* (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the GEPHI software. *PLoS One*, **9**, e98679.

Jiangxie,L.F. *et al.* (2014) Autism-associated gene Dlgap2 mutant mice demonstrate exacerbated aggressive behaviors and orbitofrontal cortex deficits. *Mol. Autism*, **5**, 32.

Kamada,T. and Kawai,S. (1989) An algorithm for drawing general undirected graphs. *Inf. Process. Lett.*, **31**, 7–15.

Kantojärvi,K. *et al.* (2011) Fine mapping of Xq11.1-q21.33 and mutation screening of RPS6KA6, ZNF711, ACSL4, DLG3, and IL1RAPL2 for autism spectrum disorders (ASD). *Autism Res.*, **4**, 228–233.

Khan,A. *et al.* (2020) An integrated systems biology and network-based approaches to identify novel biomarkers in breast cancer cell lines using gene expression data. *Interdiscip. Sci.*, **12**, 155–168.

Khoury,M. *et al.* (2012) Drawing large graphs by low-rank stress majorization. Comput. Graphics Forum, **31**, 975–984.

Kobourov,S.G. (2012) Spring embedders and force directed graph drawing algorithms. arXiv, *preprint arXiv:1201.3011*.

Koutrouli,M. *et al.* (2020) A guide to conquer the biological network era using graph theory. *Front. Bioeng. Biotechnol.*, **8**, 34.

Lafferty,A. *et al.* (2021) Molecular subtyping combined with biological pathway analyses to study regorafenib response in clinically relevant mouse models of colorectal cancer. *Clin. Cancer Res.*, **27**, 5979–5992.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Li,J. *et al.* (2014) Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Mol. Syst. Biol.*, **10**, 774.

Liu,X. *et al.* (2016) Proteomic characterization reveals that MMP-3 correlates with bronchiolitis obliterans syndrome following allogeneic hematopoietic cell and lung transplantation. *Am. J. Transplant*, **16**, 2342–2351.

Livshits,A. *et al.* (2015) Pathway-based personalized analysis of breast cancer expression data. *Mol. Oncol.*, **9**, 1471–1483.

Ma'ayan,A. (2011) Introduction to network analysis in systems biology. *Sci. Signal*, **4**, tr5.tr5.

Mallavarapu,T. *et al.* (2020) Pathway-based deep clustering for molecular subtyping of cancer. *Methods*, **173**, 24–31.

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

Muzio,G. *et al.* (2021) Biological network analysis with deep learning. *Brief. Bioinformatics*, **22**, 1515–1530.

Perozzi,B. *et al.* (2014) DeepWalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, pp. 701–710.

Pian,C. *et al.* (2021) Pathway-based personalized analysis of Pan-Cancer transcriptomic data. *Biomedicines*, **9**, 1502.

Pinto,D. *et al.* (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368–372.

Raghavan,S. *et al.* (2017) Personalized Medicine-Based approach to model patterns of chemoresistance and tumor recurrence using ovarian cancer stem cell spheroids. *Clin. Cancer Res.*, **23**, 6934–6945.

Schaefer,M.H. *et al.* (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One*, **7**, e31826.

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

Smyth,G. (2005) limma: linear models for microarray data. In: Bioinformatics Computational Biology Solutions Using R and Bioconductor. pp. 397–420.

Stark,C. *et al.* (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.*, **39**, D698–704.

Tang,J. *et al.* (2015) LINE: large-scale information network embedding. In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Florence, Italy, pp. 1067–1077.

You,Q. *et al.* (2010) Gene terrain: visual exploration of differential gene expression profiles organized in native biomolecular interaction networks. *Inf. Visual*, **9**, 1–12.

Yue,Z. *et al.* (2015) PAGER: constructing PAGs and new PAG-PAG relationships for network biology. *Bioinformatics*, **31**, i250–i257.

Yue,Z. *et al.* (2018) PAGER 2.0: an update to the pathway, annotated-list and gene-signature electronic repository for human network biology. *Nucleic Acids Res.*, **46**, D668–D676.

Zhang,F. and Chen,J.Y. (2010) Discovery of pathway biomarkers from coupled proteomics and systems biology methods. *BMC Genomics*, **11** Suppl 2, S12.

Zhang,F. and Chen,J.Y. (2013) Breast cancer subtyping from plasma proteins. *BMC Med. Genomics*, **6** Suppl 1, S6.