Parenclitic networks: uncovering new functions in biological data

Massimiliano Zanin1,2, Joaquín Medina Alcazar3, Jesus Vicente Carbajosa3, Marcela Gomez Paez3, David Papo4, Pedro Sousa1, Ernestina Menasalvas4 & Stefano Boccaletti5

1Faculdade de Ciências e Tecnologia, Departamento de Engenharia Electrotécnica, Universidade Nova de Lisboa, Lisboa, Portugal, 2Innaxis Foundation & Research Institute, José Ortega y Gasset 20, 28006, Madrid, Spain, 3Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarclón, Madrid, Spain, 4Center for Biomedical Technology, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarclón, Madrid, Spain, 5CNR- Institute of Complex Systems, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Florence, Italy.

We introduce a novel method to represent time independent, scalar data sets as complex networks. We apply our method to investigate gene expression in the response to osmotic stress of Arabidopsis thaliana. In the proposed network representation, the most important genes for the plant response turn out to be the nodes with highest centrality in appropriately reconstructed networks. We also performed a target experiment, in which the predicted genes were artificially induced one by one, and the growth of the corresponding phenotypes compared to that of the wild-type. The joint application of the network reconstruction method and of the in vivo experiments allowed identifying 15 previously unknown key genes, and provided models of their mutual relationships. This novel representation extends the use of graph theory to data sets hitherto considered outside of the realm of its application, vastly simplifying the characterization of their underlying structure.

Of the different ways of representing a multi-unit system, the one afforded by complex networks is one of the most elegant and general. Often, however, defining what can indeed be treated as a system may be highly non trivial. Suppose for instance that what one wants to study is a set of biomedical data from different individuals, e.g. various blood tests, which are in essence but a collection of scalar values without any history. Whether and how such a matter should be treated as a unitary system is not obvious. In particular, how is one to establish which entities are inside the system and which are outside its boundaries? What would the elements be of such a system and how would internal relationships among them be defined?

Prima facie, such an object study would seem to lack the physical or virtual relationships between elements of the system, which anatomic brain fibres or hyper-links respectively provide for brain tissue and pages of a web site. Nor would it appear to be possible to construct the sort of functional links that one can define when time evolving variables are associated to each node, as e.g. the time evolution of a stock price, or of brain activity in a given region.

Here, we introduce a novel way of representing as networked systems such collections of isolated, possibly heterogeneous, scalars. The final result is the creation of a network for each subject, where nodes represent features, and links are weighted according to the deviation between the values of two features and their corresponding typical relationship within a studied population. The result is what we term a parenclitic network representation, from παρέκλισις, the Greek term for “deviation”, originally used by the Greek philosopher Epicurus to designate the spontaneous and unpredictable swerving of free-falling atoms, which allows them to collide. Such a representation allows defining a system the identity of which parts and relationships (as well as the system’s boundaries) are continuously “deviated” in a context dependent manner.

The method exploits information of a set of pre-labeled subjects to unveil the presence of reference relationships between nodes. The starting point is a multi-feature description of subjects, e.g. a collection of medical measurements or gene expression levels, and their affiliation to one or multiple predefined groups. While it may be unfeasible to work with the complete data set, we consider the projection of the data into all possible planes created by pairs of features. In these planes, different methods (from simple linear correlations, up to more sophisticated data mining techniques) are used to extract a reference model for each group, accounting for the characteristics of subjects. When a new, unlabeled, subject is considered, the deviation between the associated...
Figure 1 | Schematic representation of the parenclitic network reconstruction method. (a) Graphical representation of the initial data set, composed of 20 instances (systems) and three features. Each instance is represented by a green sphere, located according to the value of its features in a 3-dimensional space. The constraint surface (gray wired surface) represents the overall standard relationship $F$ of the class. A generic unlabeled system is represented by a red sphere. (b–d) Data are then projected on each of the three possible planes. The green dashed lines represent the models extracted in each plane, i.e. $F$. The red points are the positions of the unlabeled system, and the red lines indicate the distance of the system from the models. (e) The resulting parenclitic representation is a network where nodes are associated to features, and links are weighted according to the calculated distances (coded, in this Figure, into different line widths).

data and such reference models is used to weight the link between the corresponding nodes. See Figure 1 and Methods for a more detailed description of the whole procedure.

The reader should notice that while the reconstruction method proposed here is based on the network representation technique introduced in Ref. 16, its scope has been largely widened. While the original technique only focused on linear relations between genic expression levels, here we introduce a general mathematical framework that is compatible with any type of relationship and any data set, as long as features (i.e. observables) are represented by numbers.

The topological characteristics of the resulting network can then be used to extract important information about the system. In particular, atypical conditions correspond to strongly heterogeneous networks, whereas typical or normative conditions are characterized by sparsely connected networks with homogeneous nodes16. Insofar as a network representation of each instance is constructed with reference to the population to which it is compared, this technique is by its very nature a difference seeker.

We present the results of the application of the parenclitic network representation to (i) a synthetic data set and (ii) Arabidopsis thaliana gene expression data. Of the wide range of transcriptomic analyses that have been performed in Arabidopsis, we selected a subset aimed at the characterization of gene expression responses under osmotic stress conditions. We illustrate the relevance of the proposed approach in the identification of key functional elements in gene reprogramming, and discuss how our methods compares with standard alternative methods.

**Results**

As a first step in the test of the parenclitic method’s reliability, we analyze a synthetic data set that comprises 20 instances (corresponding to sets of expression levels) and 10 features - see Methods for further details. Figure 2 reports the results obtained with this synthetic data set. The left graph depicts the behavior of features 1 and 5, for the 9 normal instances (black squares) and the abnormal one (red circle). Due to the modification of feature 5 for instance 10, the red circle deviates from the expected normal behavior (blue dashed line). The two networks, on the central and right part of the Figure, respectively represent the result of the parenclitic reconstruction technique for instances 1 and 10, i.e. for a normal and the abnormal one. Two important facts have to be highlighted. First, the network corresponding to the normal instance has a much lower link density than the abnormal one, correctly indicating that most pairwise gene relationships are close to the model prediction. Furthermore, the most central elements in the abnormal network are nodes 5 and 10, highlighting the two features that have been altered. It is worth noticing that other nodes may also have a central position, e.g. node 4, due to the noise term $\xi$ included in the data set. Overall this result indicates that the parenclitic method correctly identifies both discordance nodes.

As a second step, we used parenclitic networks to analyze gene expression of the plant Arabidopsis thaliana under osmotic stress, with the objective of identifying those genes orchestrating the plant response under this specific condition. This is of particular relevance, as abiotic stresses represent the primary cause of crop loss worldwide, lowering by more than 50% the average yields of many crop plants. Therefore, a better understanding of the mechanisms behind plant responses to such stresses, starting from the genetic level, is essential.

Expression levels have been obtained from the A1GenExpress project17, including information about the 1701 genes encompassing the transcription factor repertoire18 represented in the Arabidopsis ATH1 array used in the study at six different time points (30 min, 1 h, 3 h, 6 h, 12 h and 24 h after stress onset).

Similar data sets have been studied in the last decade by means of different techniques, e.g. co-expression networks2–4 and differential-expression analysis5–11,17. Yet, we expect the parenclitic network approach to yield complementary results. Specifically, differential-expression analyses only focus on the evolution of expression levels through time, considering each gene as independent from the others.
Co-expression networks analyze similarities between the evolutions of pairs of expression levels. Finally, the parenclitic network representation focuses on pairs of genes whose expressions depart from a reference model, thus it concentrates on differences. Furthermore, in marked contrast with classical approaches where a single network is obtained reflecting similarities across stages, in the parenclitic representation the construction of a different network for each time step allows tracking the plant response through time.

An example of the obtained networks is shown in Fig. 3 (see Methods for the details of the parenclitic representation). Namely, Fig. 3 (a) depicts the giant component of the network corresponding to 3 h after osmotic treatment. The color of links accounts for their weights, with green (red) shades indicating low (high) Z-Scores, and the size of nodes is proportional to their $\alpha$-centrality - see Methods for more details. The resulting network topologies are characterized by a highly heterogeneous structure, dominated by a small number of hubs - as can be appreciated from the zoom reported in Fig. 3 (b). Such nodes with high centrality indicate that, at 3 h, the expression levels of the corresponding genes strongly deviate from the relationships generally established at other times. This suggests that hubs are performing some specific task at this time point, and therefore that they are key actors in regulating the overall plant response to that particular stress. The parenclitic network representation allowed identifying novel candidate genes, the full list of which is reported in Table 1, that were either previously unknown or were not considered to be related to the response to osmotic stress.

To confirm these predictions, we performed an in vivo screening, in which genes corresponding to the most central nodes of each graph were artificially induced in transgenic plants, and the derived phenotype after a stress response was monitored in a typical essay by measuring the length of the root of each plant - see Methods for more details on the performed experiments. As an example, Fig. 4 reports the results obtained with seven transgenic lines, i.e. seven groups of plants in which the expression of one gene, corresponding to a parenclitic hub, was artificially induced. Specifically, Fig. 4 (a) reports the mean length of roots for the seven lines, as compared to the normal root length in the wild type (black column) grown under osmotic stress conditions. The figure clearly visualizes the fact that, in all the seven examples, induction of the corresponding gene leads to a significant functional responses in the development of the plant. The results of the in vivo screening are summarized in Fig. 5. For each of the six networks analyzed, only the 20 most central genes at each time step were considered. This figure reports the number of genes already known to be relevant for the osmotic response of the plant,
and the number of previously unknown genes disclosed by the parenclitic representation that have been successfully confirmed. Thus, the use of parenclitic network representations allowed to determine the prediction of the general form of the reference model F. Here, we have chosen the use of a simple linear regression between the expression levels of genes i and j, such that:

\[
\hat{f}_{ij} = a_i + b_i f_t^j.
\]

(1)

\(f_t^j\) being the expected value of gene j at time t, \(f_t^{ij}\) the known expression levels of gene i, and \(a_i\) and \(b_i\) two free model parameters. These two coefficients are calculated by means of a linear fit of all values corresponding to other time steps, i.e., minimizing the error of the relation:

\[
\frac{1}{n} \sum_{t=1}^{n} (\hat{f}_{ij} - f_t^{ij})^2.
\]

(2)

While more complex functions could have been used for F, the choice of a linear regression has been motivated by two considerations. First, genetic expression levels are customary transformed in order to have a linear behavior, and the calculation of linear correlations between them is a common procedure in the literature. Second, the reduced number of points available to fit the function F precludes the use of higher-order expressions, as this would result in an overfitting.

Furthermore, the reader should notice that the analysis here presented considers instantaneous interactions between genes, i.e., the value of \(f_t^j\) (at time t) is only function of \(f_t^{ij}\) and not of the historical expression of gene i. In other words, when the 24-hour expression levels are analyzed, we suppose that they are independent on the expression levels at 12 h. While this is clearly a simplification, the low temporal resolution of the available data set prevents a detailed analysis of the delayed influence of gene expressions.

The distance between the expected (corresponding to the model F) and the real value of gene j is then used to weight the link connecting nodes i and j in the network. More specifically, the weight of the link is the absolute value of the Z-Score of the distance between the expected values:

\[
| \hat{f}_{ij} - f_t^{ij} |.
\]

Synthetic data analysis. Before being applied to the Arabidopsis data set, the proposed methodology has been tested with in-silico generated information, with the additional aim of providing an additional example on how the method works. This synthetic data comprise 10 instances, each one equivalent to the set of expression levels at one time step, and 20 features. The feature (expression level) of instance i at time step t is given by the following relation:
Figure 4 | In vivo experimental verification of the predictions. (a) Mean root length corresponding to the wild type (WT, black column) and to 7 other transgenic lines in which a specific gene has been artificially induced. Whiskers represent the standard deviation corresponding to each group. Asterisks denote groups for which the distribution of root lengths is different with respect to the wild type with a 0.01 significance level. (b) Photos of one plant of each of the 8 lines, at the end of the full development process. (c) and (d) Photos of two vertical plates where plants are grown. In both cases, the left (right) photos refer to wild phenotypes (to phenotypes developed by the transgenic line).

Arabidopsis network analysis. The aim of the analysis is the identification of the most central nodes (i.e., genes) within each of the six parenclitic networks. When a node is strongly central, indeed, it is highly connected, and therefore it is part of a group of many features that deviate pairwise from the expected models.

Due to the characteristics of the network, we have opted for the $\lambda$ - centrality measure, according to which the centrality of a node is a linear combination of the centralities of those to whom it is connected. If we define a vector $\lambda$ of centralities such that its $i^{th}$ component $\lambda_i$ is the centrality of the $i^{th}$ node, we have:

$$f_i = x_i + \beta_i t + \zeta,$$

being $x_i$, $\beta_i$, and $\zeta$ random numbers drawn from a normal distribution $N(0,1)$. A strong correlation can be found between pairs of features, due to their synchronous evolution with $t$, except of the noise term $\zeta$ whose objective is to simulate the natural variability observed in genetic expression levels. Finally, features 5 and 10 of instance 10 have been incremented by 2, in order to simulate genetic expression levels that deviate from the normal behavior. The aim of this analysis is then to check whether such abnormal behavior is correctly represented in the resulting parenclitic networks.

$$f_i = x_i + \beta_i t + \zeta,$$

being $x_i$, $\beta_i$, and $\zeta$ random numbers drawn from a normal distribution $N(0,1)$. A strong correlation can be found between pairs of features, due to their synchronous evolution with $t$, except of the noise term $\zeta$ whose objective is to simulate the natural variability observed in genetic expression levels. Finally, features 5 and 10 of instance 10 have been incremented by 2, in order to simulate genetic expression levels that deviate from the normal behavior. The aim of this analysis is then to check whether such abnormal behavior is correctly represented in the resulting parenclitic networks.

$$f_i = x_i + \beta_i t + \zeta,$$

being $x_i$, $\beta_i$, and $\zeta$ random numbers drawn from a normal distribution $N(0,1)$. A strong correlation can be found between pairs of features, due to their synchronous evolution with $t$, except of the noise term $\zeta$ whose objective is to simulate the natural variability observed in genetic expression levels. Finally, features 5 and 10 of instance 10 have been incremented by 2, in order to simulate genetic expression levels that deviate from the normal behavior. The aim of this analysis is then to check whether such abnormal behavior is correctly represented in the resulting parenclitic networks.

$$f_i = x_i + \beta_i t + \zeta,$$

being $x_i$, $\beta_i$, and $\zeta$ random numbers drawn from a normal distribution $N(0,1)$. A strong correlation can be found between pairs of features, due to their synchronous evolution with $t$, except of the noise term $\zeta$ whose objective is to simulate the natural variability observed in genetic expression levels. Finally, features 5 and 10 of instance 10 have been incremented by 2, in order to simulate genetic expression levels that deviate from the normal behavior. The aim of this analysis is then to check whether such abnormal behavior is correctly represented in the resulting parenclitic networks.

$$f_i = x_i + \beta_i t + \zeta,$$

being $x_i$, $\beta_i$, and $\zeta$ random numbers drawn from a normal distribution $N(0,1)$. A strong correlation can be found between pairs of features, due to their synchronous evolution with $t$, except of the noise term $\zeta$ whose objective is to simulate the natural variability observed in genetic expression levels. Finally, features 5 and 10 of instance 10 have been incremented by 2, in order to simulate genetic expression levels that deviate from the normal behavior. The aim of this analysis is then to check whether such abnormal behavior is correctly represented in the resulting parenclitic networks.

$$f_i = x_i + \beta_i t + \zeta,$$

being $x_i$, $\beta_i$, and $\zeta$ random numbers drawn from a normal distribution $N(0,1)$. A strong correlation can be found between pairs of features, due to their synchronous evolution with $t$, except of the noise term $\zeta$ whose objective is to simulate the natural variability observed in genetic expression levels. Finally, features 5 and 10 of instance 10 have been incremented by 2, in order to simulate genetic expression levels that deviate from the normal behavior. The aim of this analysis is then to check whether such abnormal behavior is correctly represented in the resulting parenclitic networks.

$\lambda x_i = \sum_j W_{ij} (x_j + z_j) = \lambda x$, (4)

Here, $W$ is the weight matrix of the network, and $W_{ij}$ codifies the weight of the link connecting nodes $i$ and $j$. Notice that this is equivalent to an eigenvalue problem, with constant $\lambda$ defining weak connections between all the nodes of the network. In order to have meaningful results, $\lambda$ should be smaller than the spectral radius of $W$.

Osmotic stress tolerance test. For the screening of the transcription factors identified by the parenclitic model, the Arabidopsis thaliana inducible lines from Transplanta collection were used, with the ecotype Columbia (Col-0) as the Wild Type. Each one of the transgenic Arabidopsis lines of the collection expresses a single Arabidopsis transcription factor under the control of the $\beta$-stradiol inducible promoter.

For osmotic stress screening, seeds from control plants (Col-0) and at least two independent T3 homozygous transgenic lines (Transplanta collection) of each transcription factor were sterilized, vernalized for 2 days at 4°C and plated onto Petri
Best 20 identified genes by PN approach

| Time after the onset of stress treatment | Analyzed: positive | Analyzed: false positive | Already described | Others (not tested) |
|-----------------------------------------|--------------------|-------------------------|------------------|--------------------|
| 30 min                                  | 20                 | 0                       | 0                | 0                  |
| 1 h                                     | 15                 | 5                       | 0                | 0                  |
| 3 h                                     | 10                 | 10                      | 0                | 0                  |
| 6 h                                     | 5                  | 15                      | 0                | 0                  |
| 12 h                                    | 0                  | 0                       | 0                | 0                  |
| 24 h                                    | 0                  | 0                       | 0                | 0                  |

Figure 5 | Outcome of the experimental results. Bars account for the 20 most central genes at each time step. For the six time steps considered, bar colors are coded according to the following stipulations: genes previously considered not to be involved in the plant’s response to osmotic stress, that were respectively experimentally proven to develop (green) or to fail to develop (red) a statistically significant difference in the phenotype with respect to the wild-type phenotype; (cyan) genes predicted by the parenclitic analysis that were previously associated with the stress response in the Literature; and (gray) previously unknown genes, which could not be tested experimentally, due to their unavailability in the TRANSPLANTA collection.

Acknowledgments

This work was supported by grants from Instituto Nacional de Investigacion y Tecnologia Agraria y Alimentaria (INIA; project number 2012-0008-C01) and the Spanish Ministry of Science and Innovation (project number: BIO2010-14871), and the MERIT Project (FP7 ITN2010-264474). The authors also acknowledge the computational resources, facilities and assistance provided by the Centro computazionale di Ricerca sui Sistemi COmplessi (CRESCO) of the Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), and the facilities provided by CESVIMA (Spain). Finally, authors gratefully thank Shlomo Havlin for many useful discussions and suggestions.

Author contributions

M.Z. conceived and elaborated the method for parenclitic network reconstruction. J.M.A., J.V.C. and M.G.P. performed the experiment on the Arabidopsis thaliana. M.Z., D.P., P.S., E.M. and S.B. analyzed the data and prepared the figures. M.Z., J.M.A., J.V.C., D.P. and S.B. wrote the text of the Manuscript. All Authors reviewed the Manuscript.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zannin, M. et al. Parenclitic networks: uncovering new functions in biological data. Sci. Rep. 4, 5112; DOI:10.1038/srep05112 (2014).