

Appendix S1.

SUPPLEMENTARY MATERIAL AND METHODS

Auto Contractive Map insights

The Auto Contractive Map (AutoCM, Semeion[®], Rome, Italy) is a fourth-generation unsupervised artificial neural network (ANN) that can be used to assess databases, elucidating the network of links between considered variables and enabling network-based clustering of patients.

The database is regarded as a series of matrixes and weighted, it is filtered using a minimum spanning tree algorithm (MST), generating a graph that contains biological evidence which has already been tested in the medical field (S1–S4). The aim of this data-mining model is to reveal hidden trends and associations among variables. The algorithm can be used to create a semantic connectivity map in which non-linear associations are preserved and explicit connection patterns are described. This approach maps relevant connections between and among variables, and the principal hubs of the system. Hubs can be defined as variables with the maximum number of connections in the map. From a mathematical standpoint, the specificity of the AutoCM algorithm lies in its ability to minimize a complex cost function by comparison with traditional algorithms.

Traditional cost minimization function:

$$E = \text{Min} \left\{ \sum_i^N \sum_j^N \sum_q^M u_i^q \cdot u_j^q \cdot \sigma_{i,j} \right\}$$

AutoCM cost minimization function:

$$E = \text{Min} \left\{ \sum_i^N \sum_j^N \sum_k^N \sum_q^M u_i^q \cdot u_j^q \cdot u_k^q \cdot A_{i,j} \cdot A_{i,k} \right\};$$

$$A = \left(1.0 - \frac{w}{C} \right);$$

N = Number of Variables (Columns);

M = Number of Patterns (Rows).

Comparing the 2 cost functions, traditional minimization includes only second-order effects, whereas the AutoCM algorithm also considers a third-order effect. Thus, the AutoCM algorithm can reveal similarities between variables that are completely embedded in the dataset and invisible to other, classical, tools. This approach describes a context typical of living systems, in which there is continuous, time-dependent complex change in the value of the variables. AutoCM can also learn under difficult circumstances, when the connections on the main diagonal of the second connections matrix are removed, for instance. When the learning process is organized in this way, AutoCM identifies specific relationships between each variable and all the others. From an experimental point of view, the ranking of its connections matrix consequently

seems to equate to the ranking of the joint probability between each variable and the others. AutoCM requires a training phase in which the algorithm learns how variables are interconnected. The AutoCM's learning algorithm can be summarized in 4 orderly steps: (a) signal transfer from the input into the hidden layer; (b) adaptation of the connections value between the input layer and the hidden layer; (c) signal transfer from the hidden layer into the output layer; (d) adaptation of the connections value between the hidden layer and the output layer.

The MST equates to the "nervous system" of a given dataset. From the sum of all the connection strengths among all the variables, the total energy of the system is obtained. The MST selects only the connections that minimize this energy, which are the only ones needed to keep the system coherent. So, all the links included in the MST are fundamental, but not every "fundamental" link in the dataset needs to be in the MST. This limitation is intrinsic in the nature of the MST: every link that gives rise to a cycle in the graph (viz., that destroys the graph's tree-like shape) is removed, whatever its strength and meaningfulness. To fix this shortcoming, and better capture the intrinsic complexity of a dataset, more links need to be added to the MST, based on 2 criteria: (i) the new links have to be relevant in the quantitative sense; and (ii) from the qualitative standpoint, they have to be able to generate new, regular cyclic microstructures. The additional links superimposed on the MST graph generate a maximally regular graph (MRG).

This MRG is the graph with the hub function achieving the highest value among all the graphs generated by putting the connections previously skipped during the computation of the MST back into the original MST, one by one. In other words, starting from the MST, the MRG presents the largest number of regular microstructures, highlighting the most important connections of the dataset. The resulting "diamond" expresses the core complexity of the system and, in our particular case, the core of the disease.

REFERENCES

- S1. Buscema M, Grossi E. The semantic connectivity map: an adapting self-organising knowledge discovery method in data bases. Experience in gastro-oesophageal reflux disease. *Int J Data Min Bioinform* 2008; 2: 362–404.
- S2. Buscema M, Grossi E, Snowdon D, Antuono P. Auto-Contractive Maps: an artificial adaptive system for data mining. An application to Alzheimer disease. *Curr Alzheimer Res* 2008; 5: 481–498.
- S3. Buscema M. A novel adapting mapping method for emergent properties discovery in data bases: experience in medical field. *Proceeding of IEEE International Conference on Systems, Man and Cybernetics (SMC 2007)*, October 2007. IEEE, Montreal, Canada 2007: 3457–3463.
- S4. Licastro F, Porcellini E, Chiappelli M, Forti P, Buscema M, Ravaglia G, Grossi E. Multivariable network associated with cognitive decline and dementia. *Neurobiol Aging* 2010; 31: 257–269.