Efficient combination of pairwise feature networks

Pau Bellot   PAU.BELLOT@UPC.EDU
Department of Signal Theory and Communications, Technical University of Catalonia, UPC-Campus Nord, C/ Jordi Girona, 1-3, 08034, Barcelona, Spain.

Patrick E. Meyer   PATRICK.MEYER@ULG.AC.BE
Bioinformatics and Systems Biology (BioSys), Faculty of Sciences, Université de Liège (ULg), 27 Blvd du Rectorat, 4000 Liège, Belgium. *http://www.biosys.ulg.ac.be/*

Editor: Vincent Lemaire, Demian Battaglia, Isabelle Guyon, Jordi Soriano

Abstract

This paper presents a novel method for the reconstruction of a neural network connectivity using calcium fluorescence data. We introduce a fast unsupervised method to integrate different networks that reconstructs structural connectivity from neuron activity. Our method improves the state-of-the-art reconstruction method General Transfer Entropy (GTE). We are able to better eliminate indirect links, improving therefore the quality of the network via a normalization and ensemble process of GTE and three new informative features. The approach is based on a simple combination of networks, which is remarkably fast. The performance of our approach is benchmarked on simulated time series provided at the connectomics challenge and also submitted at the public competition.

Keywords: network reconstruction algorithms, elimination of indirect links, connectomes

1. Introduction

Understanding the general functioning of the brain and its learning capabilities as well as the brain structure and some of its alterations caused by disease, is a key step towards a treatment of epilepsy, Alzheimer's disease and other neuropathologies.

This could be achieved by recovering neural networks from activities. A neural network is a circuit formed by a group of connected (physically or by means of neural signals) neurons that performs a given functionality. These circuits are responsible for reflexes, senses, as well as more complex processes such as learning and memory.

Thanks to fluorescence imaging, we can easily measure the activity of a group of neurons. The changes of fluorescence recorded from the neural tissue are proved to be directly corresponding to neural activity. With calcium imaging one can study the neural activity of a population of neurons simultaneously allowing to uncover the function of neural networks.

But, recovering the exact wiring of the brain (connectome) including nearly 100 billion neurons that have on average 7000 synaptic connections to other neurons is still a daunting task. Hence, there is a growing need for fast and accurate methods able to reconstruct these networks. That is why the ChaLearn non-profit organization has proposed the connectomics challenge. The goal of the competition is to reconstruct the structure of a neural network from temporal patterns of activities of neurons.

2. Typical Methods

There is a wide variety of reconstruction algorithms that are able to infer the network structure from time series. Even though one of the least controversial necessary criterion to establish a cause-effect relationship is temporal precedence, many causal inference algorithms only require conditional independence testing Pearl (2000), or, more recently, joint distribution of pairs of variables Janzing et al. (2010). The work of Clive Granger has lead to a framework that has received a lot of attention due to its simplicity and the successful results Popescu and Guyon (2013).

2.1. Correlation with discretization

Here, we present a quick review of the simplest method to reconstruct a network, based on the correlation coefficient. The correlation is a standard method to quantify the statistical similarity between two random variables X and Y and it is defined as:

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

where μ_X and μ_Y are respectively the expected values of X and Y, and σ_X and σ_Y are the standard deviations.

Once we have a correlation coefficient between each pair of neurons, we can construct a co-activity network. If the correlation is greater than a threshold, then the neurons are connected in an undirected way, this strategy is presented at Butte and Kohane (2000), where instead of using the correlation they use Mutual Information which can be seen as a non-linear dependency measure.

2.2. Generalized Transfer Entropy

Here, we present a quick review of one of the state of the art methods, the Transfer Entropy (TE) Schreiber (2000) based measure of effective connectivity called Generalized Transfer Entropy, or GTE Stetter et al. (2012).

The basic idea behind Granger causality to test if the observations of time series of two variables A and B are due to the underlying process "A causes B" rather than "B causes A", is to fit different predictive models A (present time) and B (present time) as a function of A (past times) and B (past times). If A can be better predicted from past values of A than from past values of B, while B is also better predicted by A, then we have an indication for A being the cause of B.

Based on this idea, several methods have been derived in order to improve the results. These methods incorporate the frequency domain analysis instead of a time domain analysis Nolte and Müller (2010). One recent idea is to add contemporaneous values of B to predict A and vice versa to take into account instantaneous causal effect, due for instance to insufficient time resolution Moneta and Spirtes (2005).

Therefore, GTE can be seen as a reconstruction algorithm of causal networks based solely on pairwise interactions.

3. Our proposal: Unsupervised ensemble of CLRed pairwise features

If A activates B and this last one activates C, is very likely that co-activity networks will find a strong dependency between A and C even though the latter is an indirect link. Gene Network Inference methods have proposed different strategies to eliminate these indirect links Butte and Kohane (2000); Margolin et al. (2006); Meyer et al. (2007).

One of these strategies is the Context Likelihood or Relatedness network (CLR) method Faith et al. (2007). In order to do so, the method derives a score that is associated to the empirical distribution of the score values. Consider a score $S_{i,j}$ indicating the strength of an alleged connection between two neurons i and j. Let us call μ_i and σ_i the mean and standard deviation of this score over all neurons connected to i. The asymmetric standardized score is given as:

$$z_i = max \left(0, \frac{S_{i,j} - \mu_i}{\sigma_i} \right) \tag{2}$$

Finally, the symmetrized score is given by: $z_{ij} = \sqrt{z_i^2 + z_j^2}$. This method has a complexity of $O(n^2)$, n being the number of neurons, and requires a symmetric matrix.

Our unsupervised ensemble of pairwise features uses the CLR algorithm to eliminate indirect links and normalize the network before assembling the different CLRed pairwise features. With the second step we are able to eliminate more indirect links that are still present at one reconstructed network but not at the others. This idea comes from modENCODE Consortium et al. (2010); Marbach et al. (2012), where the authors propose an algorithm to integrate different network inference methods to construct a community networks which is capable of stabilizing the results and recover a good network. Their state-of-the-art method to combine networks is based on rank averaging. The individual ranks of each link are added together to compute the final rank. Then, the final list is computed sorting these score decreasingly. This method is also known as Ranksum, and will be referred as RS in the paper.

Instead of this procedure, our proposal that will be referred as $CLRsum$ or CS is formulated as follows:

$$CS := \sum_i^N CLR \left(feature_i \right) \tag{3}$$

A description of the workflow of our network reconstruction process is available in Figure 1 in the Supplementary Material Bellot and Meyer (2014). In this case, we have used four features that are defined in the following subsections.

3.1. Feature 1: Symmetrized GTE

The first pairwise measure is a modification of the state-of-the-art method GTE (see section 2.2), since we apply the CLR method the recovered network should be undirected. Indeed, the GTE method provides a non-symmetric score ($gte_{i \to j} \neq gte_{j \to i}$), we symmetrize it by taking the most conservative score recovered by GTE. This symmetrized GTE network is denoted as GTE_{sym}, and is defined as follows: $gte_{i,j} = min(gte_{i \to j}, gte_{j \to i})$.

3.2. Feature 2: Correlation of the extrema of the signals

The second pairwise measure is based on the correlation of both signals ($corr(X_i, X_j)$). But, doing so we are not able to discriminate between true regulations and indirect effects or light scattering effects. We observed experimentally with the provided networks and their respective ground-truth, that the correlation between the signals when one of both neurons is spiking is statistically more informative than the plain correlation.

The quantile $q_{k\%}(x)$ is the data value of x where we have k % of the values of x above it. The higher the quantile the stronger the statistical correlation between the measure and the connectome network. However, in order to be able to recover a non-spurious correlation at least several hundreds of samples are required. First, we capture the quantile $\alpha\%$ of both signals, and compute the correlation using only the points of both signals that are above the quantile:

$$\text{Let } t_k := X_i(t) \geq q_{\alpha\%}(X_i) \text{ and } t_l := X_j(t) \geq q_{\alpha\%}(X_j)$$
$$ct_{i,j} = corr\left(X_i(t_n), X_j(t_n)\right) \text{ with } t_n := t_k \cup t_l \tag{4}$$

Computing previous equation between each pair of different neurons we obtain the $CT_{\alpha\%}$ network.

3.3. Feature 3: mean squared error of difference signal

The third pairwise feature that we have found experimentally, is complementary to feature 2. Instead of computing the correlation on the spikes, this feature uses the mean squared error of the points where the two signals disagree the most (once both have been normalized by a centering and scaling). The normalization process is defined as $X_i^s := (X_i - \mu_{X_i})/\sigma_{X_i}$.

First, we compute the difference between the two scaled different signals ($i \neq j$) and keep the points where they differ the most. To get such particular time points, we also rely on an small quantile $\alpha\%$:

$$\text{Let } f_{i,j}(t) = X_i^s(t) - X_j^s(t) \text{ and } t_k := d_{i,j}(t) \geq q_{\alpha\%}(f_{i,j})$$
$$\text{Then } X_i' := X_i^s(t_k) \; X_j' := X_j^s(t_k) \tag{5}$$

Once the points of interest (t_k) are extracted, the mean square error between $p_{i,j} := X_i' - X_j'$ and $p_{j,i} := X_j' - X_i'$ is computed. This leads to a non-symmetrical measure. As explained before, CLR requires a symmetric matrix. In order to symmetrize the matrix we take the minimum of $p_{i,j}^2$ and $p_{j,i}^2$ as has been done in feature 1: $cd_{i,j} = \min\left((X_i' - X_j')^2, (X_j' - X_i')^2\right)$.

This measure is computed between each pair of different neurons to obtain the $MD_{\alpha\%}$ network.

3.4. Feature 4: range of difference signal

The last pairwise measure that we have found correlated to the connectome is the range of the difference between two neuron signals.

For every pair of neurons we compute the difference between the two different signals ($i \neq j$): $df_{i,j} := X_i - X_j$

Then, the measure captures the range of $df_{i,j}$, but in order to be robust to the presence of noise the range is not the difference between the largest and smallest values of $df_{i,j}$,

but the average over the 10th maximal/minimal values of $df_{i,j}$. This measure is computed between each pair of different neurons to obtain the network RD. In order to obtain a similarity network, we invert the network as follows:

$$RD := max(R) - R \text{ with } diag(RD) = 0 \tag{6}$$

4. Experiments

The performance of our algorithm is benchmarked on the data provided at the ChaLearn connectomics challenge. The data reproduces the dynamic behavior of real networks of cultured neurons. The simulator also includes the typical real defects of the calcium fluorescence technology: limited time resolution and light scattering artifacts (the activity of given neuron influences the measurements of nearby neurons) Stetter et al. (2012).

The challenge provides different datasets that have distinct properties, we will use the datasets where the network structure is also provided, i.e, 10 big datasets of 1000 neurons and 5 small datasets of 100 neurons. The network inference problem can be seen as a binary decision problem: after the thresholding of the network provided by the algorithm, the final decision can be seen as a classification: for each possible pair of neurons, the algorithm either define a connection or not. Therefore, the performance evaluation can be assessed with the usual metrics of machine learning like Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves. The ChaLearn Connectomics proposes as a global metric the use of the area under the ROC curve (AUC), however, ROC curves can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution, as typically encountered in sparse network problems. To tackle this problem, precision-recall (PR) curves have been proposed as an alternative to ROC curves Sabes and Jordan (1995). For this reason, we present in Table 1 the Area Under PR curve (AUPR) and compare our method with GTE and the state-of-the-art combination of these features *Ranksum*. The results of GTE are obtained with the software available online dherkova (2014) using as conditioning levels the values $\{0.05, 0.10\}$ for the iNet1-Size100-CC's networks and $\{0.15, 0.20\}$ for the big datasets. We have also computed a statistical test to discard non-significant results. First, we compute the contribution of each link to the area under the curve and then we apply the Wilcoxon test on the resulting vectors Hollander et al. (2013). If the best result of each dataset have a p-value smaller than 5% it is typed in italic font and boldfaced.

Table 1 shows the performance of our individual networks ($CT_{0.1\%}$, $MD_{0.1\%}$, RD) and we can observe that it depends on the properties of the network (high/low-connectivity or high/low-activity).

We also compare our community based approach with the state-of-the-art Ranksum approach, which is shown at the last column. Note that the Ranksum makes use of the original pairwise derived networks and our method used the symmetrized GTE (denoted as I^*). We can observe that our normalization and simple combination is able to improve the quality of the individual recovered networks and also improves the state-of-the-art community Ranksum. As shown in the table, our approach is competitive even though our method does not recover a directed network as GTE does. It is worth noting that using AUC as a metric we obtain similar conclusions. The table with AUC results is available in the Supplementary Material Bellot and Meyer (2014).

| | I | II | III | IV | V | | |
	GTE	corr	$CT_{0.1\%}$	$RD_{0.1\%}$	$MD_{0.1\%}$	$CS\left(I^*, III, VI, V\right)$	$RS\left(I, III, VI, V\right)$
highcc	0.163	0.051	0.088	0.166	0.125	*0.330*	0.184
highcon	0.199	0.030	0.120	0.125	0.073	*0.241*	0.184
iNet1-Size100-CC01inh	*0.242*	0.117	0.117	0.106	0.123	0.180	0.158
iNet1-Size100-CC02inh	*0.247*	0.113	0.150	0.103	0.120	0.223	0.181
iNet1-Size100-CC03inh	*0.333*	0.116	0.198	0.130	0.131	0.314	0.237
iNet1-Size100-CC04inh	*0.398*	0.120	0.206	0.187	0.158	0.394	0.297
iNet1-Size100-CC05inh	0.366	0.120	0.208	0.288	0.179	*0.423*	0.366
iNet1-Size100-CC06inh	0.538	0.204	0.188	0.371	0.318	*0.582*	0.480
lowcc	0.085	0.015	0.085	0.031	0.022	*0.126*	0.083
lowcon	0.093	0.023	0.025	0.024	0.031	*0.196*	0.125
normal-1	0.164	0.028	0.085	0.110	0.061	*0.251*	0.155
normal-2	0.169	0.028	0.105	0.095	0.048	*0.242*	0.153
normal-3-highrate	0.201	0.057	0.037	0.143	0.073	*0.293*	0.181
normal-3	0.193	0.025	0.098	0.094	0.044	*0.260*	0.150
normal-4-lownoise	0.141	0.030	0.120	0.086	0.053	*0.271*	0.156
normal-4	0.139	0.026	0.085	0.082	0.046	*0.254*	0.140
Mean	0.229	0.069	0.120	0.134	0.100	**0.286**	0.202

Table 1: Area Under Precision Recall scores for each inference method at the datasets of the connectomics challenge (the higher the better). The Ranksum makes use of the original pairwise inferred networks while our method use the symmetrized GTE (denoted as I^*). The best statistically significant results tested with a Wilcoxon test are highlighted.

Additionally to the results shown at table 1 we also have used our method in the test and validation networks where the network is unknown. Using the connectomics submission tool we obtain 0.90402 score of Area under the ROC curve, and we would have been ranked in the 30th position.

As stated previously the big advantage of our method is the low complexity $O(n^2)$. The CPU time needed to compute the different features for the big datasets in a 2 x Intel Xeon E5 2670 8C (2.6 GHz), has a mean of 1282.86 minutes for the GTE, 3.31 minutes for $CT_{0.1\%}$, 66.53 minutes for $MD_{0.1\%}$ and 21.17 minutes for RD. The process of CLRsum is almost instantaneous once we have the individual features, and therefore the computation time is the sum of the time needed to compute the individual features. Hence, our proposal improves GTE with a negligible overload of time.

5. Conclusion

An unsupervised network inference method for neural connectomics has been presented. This method improves the state-of-the-art network inference method GTE relying on $CLRsum$ consensus among GTE and three new informative features.

We have compared our method experimentally to two state-of-the-art network inference methods, namely GTE and correlation network, on the connectomics challenge datasets. The experimental results showed that our proposal is competitive with state-of-the-art algorithms.

Acknowledgments

References

Pau Bellot and Patrick E Meyer. Efficient combination of pairswise feature networks - supplemental material. http://bellot.geek-lab.org/publications/efficient-combination-of-pairwise-feature-networks, 2014. Accessed: 2014-07-22.

Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pac Symp Biocomput*, volume 5, pages 418–429, 2000.

dherkova. Transfer entropy - causality. https://github.com/dherkova/TE-Causality/tree/connectomics_challenge, 2014. Accessed: 2014-06-01.

Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.

Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

Dominik Janzing, Patrik O. Hoyer, and Bernhard Schölkopf. Telling cause from effect based on high-dimensional observations, 2010.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

Patrick E Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, page 79879, 2007.

The modENCODE Consortium et al. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797, 2010.

Alessio Moneta and Peter Spirtes. Graph-based search procedure for vector autoregressive models. Technical report, LEM Working Paper Series, 2005.

Guido Nolte and Klaus-Robert Müller. Localizing and estimating causal relations of interacting brain rhythms. *Frontiers in human neuroscience*, 4, 2010.

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

Florin Popescu and Isabelle Guyon, editors. *Causality in Time Series: Challenges in Machine Learning*, volume 5. Microtome, 2013.

Philip N Sabes and Michael I Jordan. Advances in neural information processing systems. In *In G. Tesauro & D. Touretzky & T. Leed (Eds.), Advances in Neural Information Processing Systems*. Citeseer, 1995.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Olav Stetter, Demian Battaglia, Jordi Soriano, and Theo Geisel. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Comput Biol*, 8 (8):e1002653, 08 2012.