Supplementary Materials

for manuscript

MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data

Nadir Sella [1,2], Louis Verny [1,2], Guido Uguzzoni [1,2], Séverine Affeldt [1,2,3] and Hervé Isambert [1,2*]

[1]Institut Curie, PSL Research University, CNRS, UMR168, 26 rue dUlm, 75005 Paris, France,
[2]Sorbonne Universités, UPMC Univ Paris 06, 4, Place Jussieu, 75005 Paris, France and
[3] Current address: LIPADE, University of Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France.

1 User Documentation

A Tutorial and detailed User Guide documentations for `MIIC online` server are available at https://miic.curie.fr

2 `MIIC online` pipeline

The work-flow of `MIIC online` server is outlined in Fig. S1.

It consists of *i)* an input layer including the data, default or user-defined parameters and optional supplementary files uploaded by the user, *ii)* the algorithmic core of the network reconstruction and *iii)* an output layer with all interactive visualizations and analyses about the results.

`MIIC` algorithmic core *(ii)* includes three main steps detailed in the Methodological Sections of [Verny *et al.*, 2017].

These three algorithmic steps are summarized below:

Step1: Learning the network 'skeleton'

Starting from a fully connected undirected graph, `MIIC` iteratively prunes the edges that are not required to account for the observed correlations in the available data, as the corresponding correlations can already been explained by indirect paths without the need for additional edges between some of the nodes. `MIIC` algorithm proceeds as follows based on information-theoretic results [Affeldt *et al.*, 2015, Affeldt *et al.*, 2016].

Given two nodes X and Y, `MIIC` looks for the most significant contributors susceptible to explain the mutual information between X and Y, $I(X;Y)$, and iteratively removes their contributions as,

$$I(X;Y|\{A_i\}) = I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1)... - I(X;Y;A_i|\{A_{i-1}\}) \qquad (1)$$

until the residual conditional mutual information between X and Y given $\{A_i\}$, $I(X;Y|\{A_i\})$, becomes lower than the associated complexity loss of the graphical model without the XY edge, $k_{X;Y|\{A_i\}}/N$, where N is the number of independent samples. Otherwise, if $I(X;Y|\{A_i\}) > k_{X;Y|\{A_i\}}/N$, the XY edge is retained, if no additional contributor can be found to account for the residual conditional mutual information between X and Y. This first step of `MIIC` algorithm returns an undirected graph, referred to as the network 'skeleton'.

Step2: Edge filtering (optional) based on Confidence ratio

The edge filtering step (optional) allows to remove additional edges from the first network skeleton obtained in Step 1, according to an edge-specific confidence assessment [Verny *et al.*, 2017]. It is based on the probability to delete the edge XY between nodes X and Y, which can be estimated as,

$$P_{XY} = e^{-NI'(X;Y|\{A_i\})} \tag{2}$$

where N is the number of independent samples in the data and $I'(X;Y|\{A_i\}) = I(X;Y|\{A_i\}) - k_{X;Y|\{A_i\}}/N$.

The probability P_{XY} is then evaluated for each retained edge of the skeleton obtained using the actual dataset *versus* multiple randomized instances of the same dataset. This allows to compute the following edge-specific confidence ratio:

$$C_{XY} = \frac{P_{XY}}{\langle P_{XY}^{\text{rand}} \rangle} \tag{3}$$

where $\langle P_{XY}^{\text{rand}} \rangle$ is the mean probability to remove the edge XY averaged over multiple randomized datasets. Hence, a smaller confidence ratio, C_{XY}, implies a higher statistical confidence on the retained XY edge.

Step3: Edge orientation

Finally, given the skeleton obtained in Step 1, possibly filtered in Step 2, `MIIC` infers edge orientations based on the signature of causality in observational data as follows [Verny *et al.*, 2017].

First, `MIIC` sorts unshielded triples, *i.e.* $X - Z - Y$ without XY edge, by decreasing absolute value of their three-point conditional mutual information including finite size complexity correction, $|I'(X;Y;Z|\{A_i\})|$, where $\{A_i\}$ is the (possibly empty) set of contributors accounting for the removal of the XY edge, *i.e.* with $I'(X;Y|\{A_i\}) < 0$. Then, `MIIC` orients the XZ and/or ZY edges as,

• If $I'(X;Y;Z|\{A_i\}) < 0$, it forms a V-structure: $X{*\!\!\rightarrow} Z \leftarrow\!\!{*}Y$

• If $I'(X;Y;Z|\{A_i\}) > 0$ and $X{*\!\!\rightarrow} Z$, the second edge is oriented as to form a non-v-structure: $X{*\!\!\rightarrow} Z \rightarrow Y$

where the endpoint mark $*$ stands for either an arrow head $>$ or tail $-$. This enables, in particular, to obtain bidirected edges, *e.g.* $Z \leftrightarrow Y$, shared by two v-structures, *e.g.* $X{*\!\!\rightarrow} Z \leftrightarrow Y \leftarrow\!\!{*}W$, which reflects the presence of unobserved (latent) causes such as, $Z \dashleftarrow L \dashrightarrow Y$.

3 Examples of causal *versus* non-causal networks

In this section we illustrate the use of `MIIC online` server with two examples of network reconstruction from real biological data.

The first example is an inherently causal network corresponding to directed regulatory interactions between specific transcription factors involved in hematopoietic stem cell differentiation, while the second example is a non-causal network corresponding to undirected symmetric physical interactions between amino acid residues in close contact in a protein structure.

As discussed in the main text, these different types of causal and non-causal networks cannot be reconstructed with the single existing online server, which are all designed to learn specific classes of directed *or* undirected networks without the possibility to compare between alternative classes of graphical models. This prevents all existing network reconstruction servers to uncover or rule out causality in observational data.

By contrast, MIIC online does not require the user to select *a priori* the type of causal or non-causal underlying model, as MIIC algorithm learns the most appropriate causal, non-causal or mixed model given the available data.

3.1 Reconstruction of regulatory networks from single cell expression data

This first example concerns the reconstruction of blood stem cell regulatory network models from single-cell molecular profiles. The mammalian blood system is maintained throughout the adult lifetime by hematopoietic stem cells (HSCs) that differentiate into all mature blood cell types. Differentiation of HSCs toward alternative lineages is controlled by transcription factors within organized regulatory programs that can be modeled as transcriptional regulatory networks.

While hematopoiesis in adult has been extensively studied and well-characterized at cell population level, cell fate decisions are in fact made at the level of individual cells and lead to heterogeneous cell populations. Recent developments of high-throughput single-cell technologies, such as quantitative real-time PCR (qRT-PCR) and RNA sequencing (RNA-Seq), now provide unique tools to study such differentiation processes and corresponding regulatory networks at single-cell level.

In this section, we analyze the recent dataset obtained by [Hamey *et al.*, 2017], which contains qRT-PCR gene expression profiles of 48 genes including 34 transcription factors for 2,167 single HSCs and progenitor cells.

The input dataset used for MIIC online reconstruction includes all 34 transcription factors and has been discretized into binary levels corresponding to expressed *versus* non-expressed genes, as suggested by the clearly bimodal distributions of qRT-PCR expression profiles. As expected, no significant correlation bias between successive single cell samples is identified with MIIC online correlation analysis. Hence, all 2,167 single cell expression profiles can be considered as independent samples for the network reconstruction.

The network inferred by MIIC online is displayed in Fig. 1 of the main text (zoomed view) and Fig. S2 (full network). The edges in the reconstructed network have been filtered using a confidence ratio threshold of 10^{-1} (Step 2 of MIIC algorithm) and their width reflects their estimated confidence. They represent direct regulatory interactions between regulator and target transcription factors. In particular, we observe that nearly all predicted edges are directed, as expected for transcriptional regulatory networks, with red edges indicating gene activation and blue edges indicating gene repression regulations.

MIIC predicted network corresponds to a global transcriptional regulatory network, as it combines expression profiles of HSCs with different progenitor cell types [Hamey *et al.*, 2017]. This network exhibits a number of known central regulators such as *MECOM*|*EVI1*, *GATA1* and *GATA2*, with regulatory interactions documented in the literature, such as *MECOM* → *PBX1* [Yuan *et al.*, 2015], *MECOM* → *GATA2* [Yuasa *et al.*, 2005] and *GATA2* → *TAL1*|*SCL* [Chan *et al.*, 2006].

3.2 Reconstruction of residue-residue interaction network in protein structure from homolog genomic sequences

The three-dimensional structure similarity between homologous proteins imposes strong constraints on their sequence variability. This gives rise to correlated substitution patterns among amino acid residues at different sequence positions of a protein family. It has long been suggested that these correlations can be exploited to infer spatial contacts within the tertiary protein structure [Altschuh*et al.*, 1987][Neher, 1994]. In the last years several methods have been proposed to disentangle direct and indirect correlations, that represents one of the major difficulties for the success of the approach [Burger *et al.*, 2008] [Weigt *et al.*, 2009] [Morcos *et al.*, 2011] [Marks *et al.*, 2011].

In this section, we show the efficacy of MIIC algorithm to retrieve the internal protein contact network for a widely studied protein family: the *response regulator receiver domain* (Pfam code PF00072). This extremely abundant protein family is involved in bacterial signal transduction and acts as a transcription factor interacting with specific DNA binding domains. This family is especially suited to assess the performance of inference methods for protein contact network as (1) it contains a great number of sequenced proteins (63,624), (2) several protein structures belonging to this family have been experimentally resolved, and (3) it is a classical example that has already been studied in depth in the literature [Weigt *et al.*, 2009], [Uguzzoni *et al.*, 2017].

The input dataset consists of a multiple sequence aligment (MSA) including 112 positions of the homologous sequences, which can be downloaded from the Pfam database [Bateman *et al.*, 2004]. When the whole dataset including the 63,624 homologous sequences is used as input file on MIIC online server, a warning message appears in the Result page to indicate significant correlations between samples, which do not simply decay exponentially between successive sequences in the MSA. These correlations have been discussed in the literature and are due to the phylogeny, multiple-strain sequencing, and a biased selection of sequenced species. To overcome this issue, we have used a standard procedure to reduce the redundancy due to sequence bias [Morcos *et al.*, 2011]. Namely, we filtered the MSA by randomly selecting sequences that differ from each other for at least 30% of their positions and removing the other sequences from the MSA. After this preprocessing of the data, the resulting filtered MSA contains 12,533 sequences.

The results of MIIC network prediction are presented in Figs. S3 & S4. The edges in the reconstructed network represent the residue-residue physical proximity in the 3D structure. Using Pymol [DeLano *et al.*, 2017], we can visualize the contact predictions and overlay them to available crystallographic structures.

In Fig. S3, we report the contact predictions mapped on an experimentally resolved structure (*1nxs*) downloaded from the PDB database [Berman *et al.*, 1999]. Note that MIIC predictions provide an accurate description of the contact map of the protein (green edges). Quite remarkably, we also observe that MIIC does not predict any directed edges despites its lack of *a priori* restriction on the class of (undirected, directed or mixed) reconstructed network; this prevalence of undirected edges is in fact expected from the symmetry of the physical contacts between amino acid residues, by contrast to the asymmetric regulator-target gene relationships in the transcriptional regulation network described above. In addition, we found that most false positive contacts (red edges in left panel and red dots in right panel) are actually very close to true contacts in the *1nxs* protein structure (black dots in right panel) and are related to the intrinsic heterogeneity of the different protein structures within this large family. This is clearly apparent in Fig. S4, where MIIC predictions are compared to the union of 11 contact maps of homologous protein structures, see Fig. S4 caption. As a result, most of these apparently false positive contacts in the *1nxs* protein structure turn out to be true positive contacts once the structure heterogeneity of this large protein family is taken into account.

Finally, when these results are compared with the state-of-the-art method for protein contact prediction, PlmDCA [Ekeberg *et al.*, 2013], we find that MIIC predicts a similar list of contacts and achieves similar performance as PlmDCA, as shown in Fig. S4 and Fig. S5 (upper panel). However, it is important to stress that MIIC predicts a finite list of 179 contacts, while PlmDCA sorts all potential pairwise contacts using a rank but without predicting an explicit cutoff to distinguish between actual contacts and non-contacts. Note, also, that contacts involving residues closer than 5 AA along the sequence are not displayed in Figs. S3-S4 & S5, as they correspond to ‘trivial’ contacts and are possibly affected by small gap statistics in the MSA [Feinauer *et al.*, 2014]. Hence, Figs. S3 & S4 display in fact 75 long-distance contacts out of the 179 contacts predicted by MIIC and the first 75 potential long-distance contacts inferred by PlmDCA. Interestingly, most of remaining long-distance false positive contacts, predicted by the two methods in Fig. S4, have been shown to correspond to intermolecular contacts across homodimers rather than intramolecular contacts within a single protein domain as reported in [Uguzzoni *et al.*, 2017]. Hence, while these predicted contacts are not in the tertiary structure, they nonetheless correspond to real coevolutionary signals in the MSA due to direct physical interactions between individual monomers in the quaternary assembly of the protein homodimers.

To further assess the performance of MIIC on protein contact map predictions, we have analyzed two additional protein families containing fewer homologous sequences. These are the *1a3a:a* PDB protein structure with a total of 31,922 homologous sequences and the *1mb6:a* PDB protein structure with a total of 246 sequences.

We apply the same filtering procedure as for the *response regulator receiver domain* (*1nxs*) above. This amounts to filtering sequences with more than 70% identity to reduce phylogenetic or other sampling biases, which leads to significantly reduced datasets of only 2,897 out of 31,922 sequences for the *1a3a:a* structure and only 53 out of 246 sequences for the *1mb6:a* structure.

Comparisons of MIIC and pmlDCA ranked predictions of protein map contacts are presented in Fig. S5 and show a lower accuracy of MIIC with respect to pmlDCA for these two datasets containing fewer homologous sequences. Yet, we note that, unlike MIIC, plmDCA uses the complete homologous sequence datasets through a weighting scheme of similar sequences to compensate for phylogenetic or other sampling biases. By contrast, as noted earlier, MIIC has the useful feature of providing a finite number of (mostly correct) predictions, while plmDCA provides a ranked list of predictions including essentially all possible pairs without clear cut-off, Fig. S5.

References

[Affeldt *et al.*, 2015] Affeldt S, Isambert H (2015) Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence* (UAI), 42-51.

[Affeldt *et al.*, 2016] Affeldt S, Verny L, Isambert H. (2016) 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics, *BMC Bioinformatics* **17** (Suppl 2), 12.

[Altschuh*et al.*, 1987] D Altschuh, AM Lesk, AC Bloomer, A Klug. (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol* **193**(4), 693-707.

[Bateman *et al.*, 2004] Bateman, Alex, et al. (2004) The Pfam protein families database. *Nucl Acids Res* 32.suppl: D138-D141.

[Berman *et al.*, 1999] Berman, Helen M., et al. The Protein Data Bank, 1999-. International Tables for Crystallography Volume F: Crystallography of biological macromolecules. Springer Netherlands, 2006. 675-684.

[Burger *et al.*, 2008] Burger L, Van Nimwegen E (2008) Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* **4**(1), 165.

[Chan *et al.*, 2006] Chan WYI, Follows GA, Lacaud G, Pimanda JE, Landry JR, Kinston S, et al. (2006) The paralogous hematopoietic regulators lyl1 and scl are coregulated by ets and gata factors, but lyl1 cannot rescue the early scl–/– phenotype. *Blood* **109**(5):1908-1916.

[DeLano *et al.*, 2017] DeLano, Warren L. (2002) The PyMOL molecular graphics system. http://pymol.org.

[Ekeberg *et al.*, 2013] Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E*, **87**(1), 012707.

[Feinauer *et al.*, 2014] Feinauer, C *et al.* (2014). Improving contact prediction along three dimensions. *PLoS Comput. Biol.* , **10**(10), e1003847.

[Hamey *et al.*, 2017] Hamey FK, *et al.* (2017) Reconstructing blood stem cell regulatory network models from single-cell molecular profiles, *Proc Natl Acad Sci USA* **114**(3), 5822-5829.

[Marks *et al.*, 2011] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, **6**(12), e28766.

[Morcos *et al.*, 2011] Morcos F, *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* **108**(49), E1293-E1301.

[Neher, 1994] Neher E (1994) How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA*, **91**(1), 98-102.

[Uguzzoni *et al.*, 2017] Uguzzoni G, *et al.* (2017) Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc Natl Acad Sci USA* **114**(13), E2662-E2671.

[Verny *et al.*, 2017] Verny L, Sella N, Affeldt S, Singh PP, Isambert H. (2017) Learning causal networks with latent variables from multivariate information in genomic data, *PLoS Comput Biol*, 13(10):e1005662.

[Weigt *et al.*, 2009] Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., Hwa, T. (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci USA*, **106**(1), 67-72.

[Yuan *et al.*, 2015] Yuan, X., Wang, X., Bi, K., Jiang, G. (2015). The role of EVI-1 in normal hematopoiesis and myeloid malignancies (Review). *International Journal of Oncology*, **47**, 2028-2036.

[Yuasa *et al.*, 2005] Yuasa, H *et al.* (2005). Oncogenic transcription factor Evi1 regulates hematopoietic stem cell proliferation through GATA-2 expression. *The EMBO Journal*, **24**(11), 1976-1987.

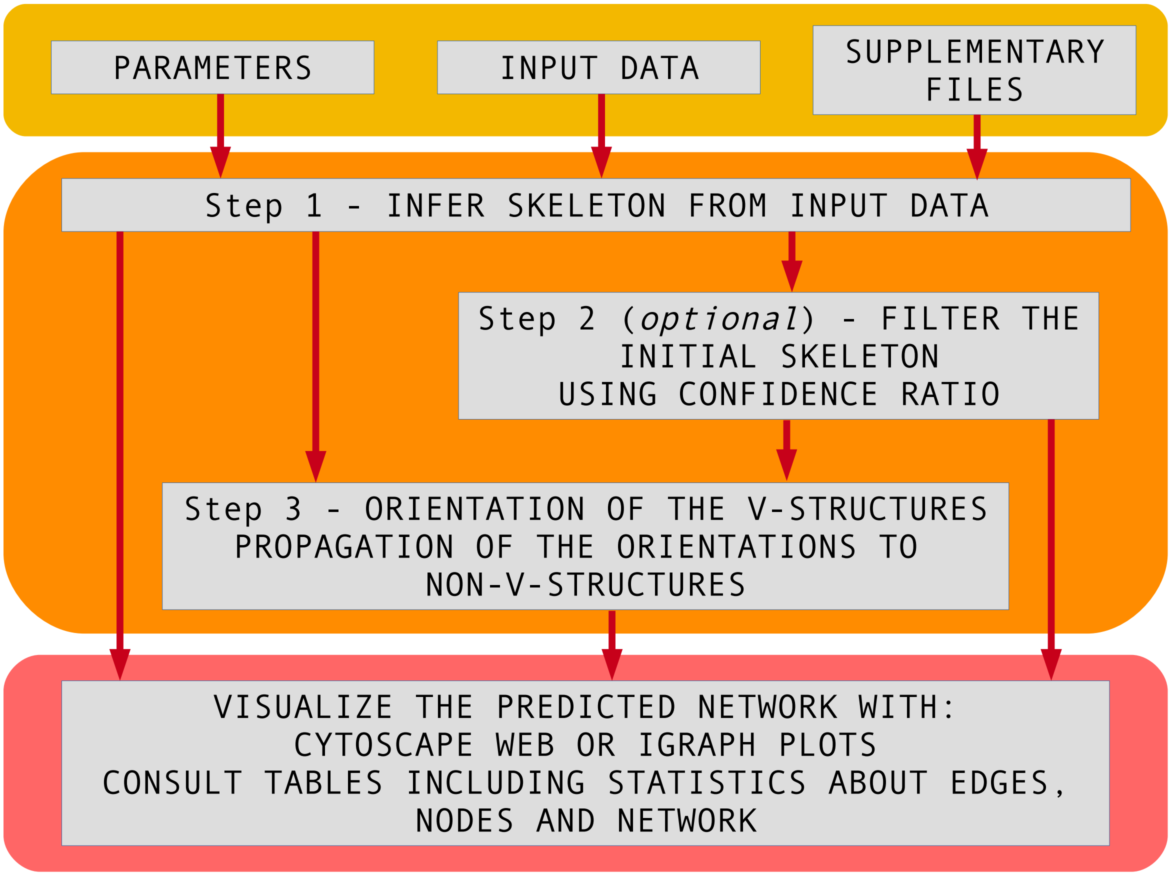
PARAMETERS
INPUT DATA
SUPPLEMENTARY FILES
Step 1 - INFER SKELETON FROM INPUT DATA
Step 2 (*optional*) - FILTER THE INITIAL SKELETON USING CONFIDENCE RATIO
Step 3 - ORIENTATION OF THE V-STRUCTURES PROPAGATION OF THE ORIENTATIONS TO NON-V-STRUCTURES
VISUALIZE THE PREDICTED NETWORK WITH: CYTOSCAPE WEB OR IGRAPH PLOTS CONSULT TABLES INCLUDING STATISTICS ABOUT EDGES, NODES AND NETWORK

Figure S1: `MIIC online` server workflow. It consists of *i)* an input layer including the data, default or user-defined parameters and optional supplementary files uploaded by the user, *ii)* the algorithmic core of the network reconstruction including three main steps and *iii)* an output layer with all interactive visualizations and analyses about the results.

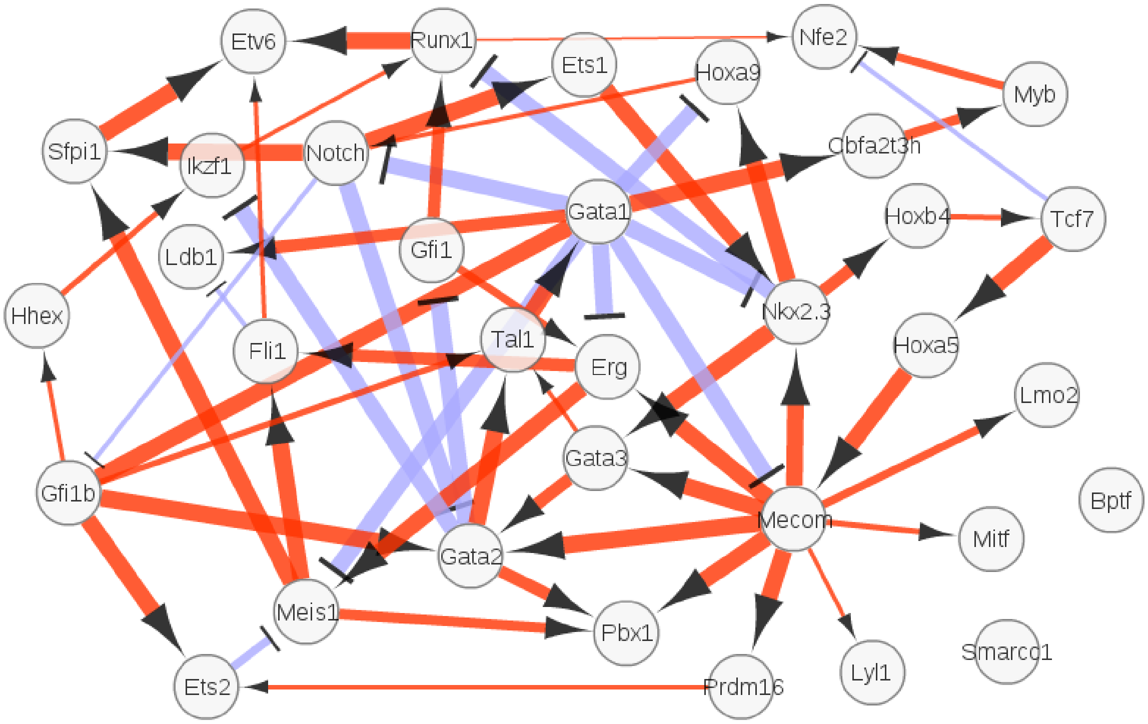

Figure S2: `MIIC online` reconstruction of the regulatory network of hematopoietic stem cell differentiation from single-cell expression data taken from [Hamey *et al.*, 2017]. See main text and Supplementary Materials for detailed information.

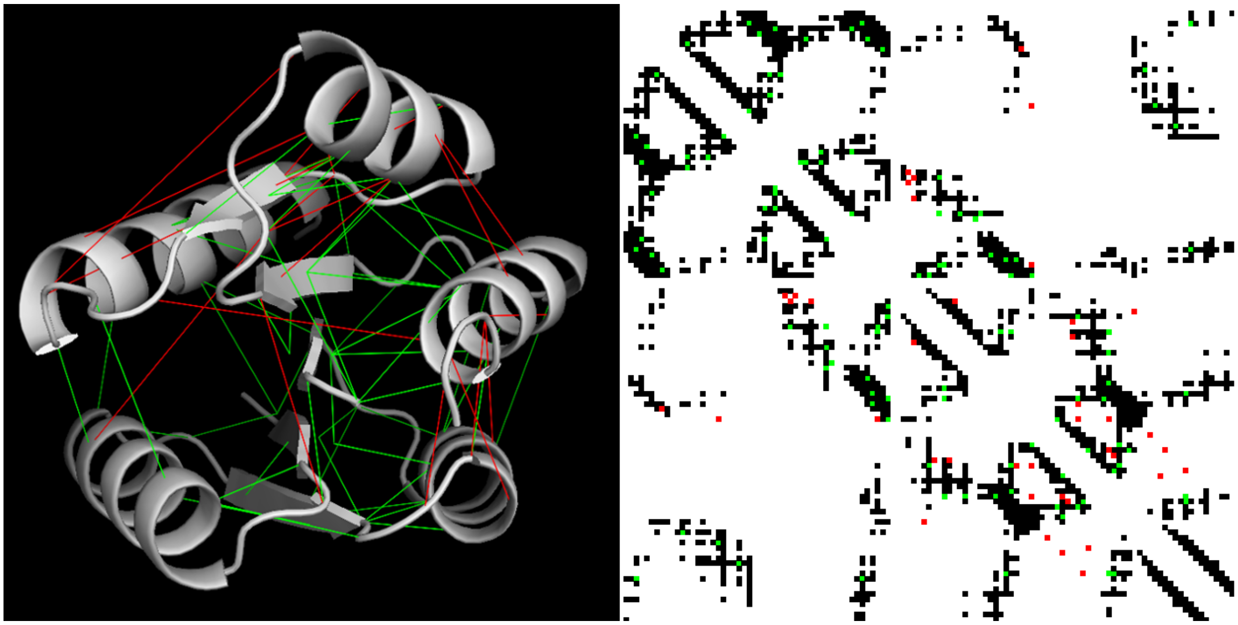

Figure S3: MIIC residue-residue contact predictions of the response regulator receiver domain (PF00072) mapped on an experimentally resolved structure (*1nxs* PDB). Contacts are defined as residues with a proximity of less than 8Å. Left panel: protein 3D structure with correct predictions in green and apparent errors in red, see however Fig. S4. Right panel: 2D contact map with experimental contacts in black and predictions with same color code as in the left panel.

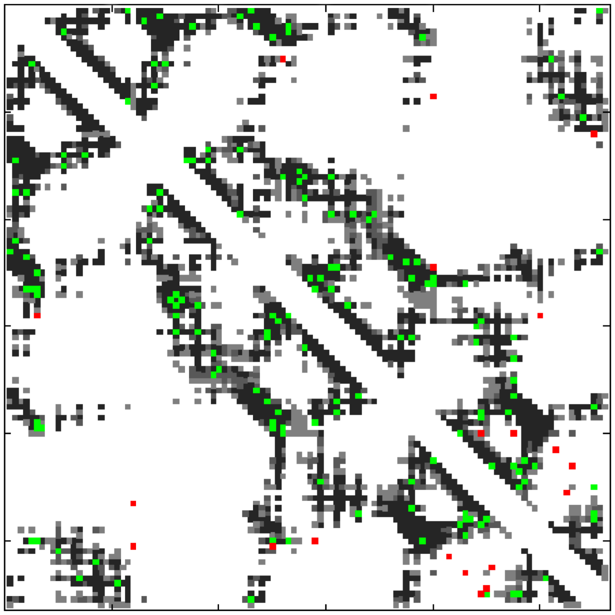

Figure S4: Contact map predictions of `MIIC` (upper triangular region) and plmDCA (lower triangular region) compared with the union of 11 experimental contact maps (from the following PDB structures: *1nxs, 1zes, 2pln, 2zwm, 3nnn, 3r0j, 2rdm, 6chy, 1l5y, 2vuh, 4l4u*). Structural contacts are displayed in black (if shared in all 11 models) or gray (if present in at least one of the 11 structures), while correct and erroneous predictions are shown in green and red, respectively. Note that the two methods present only small differences in the number of correct and erroneous predictions. Besides, many of the apparently erroneous contact predictions are in fact due to intermolecular interactions across the protein homodimers [Uguzzoni *et al.*, 2017].

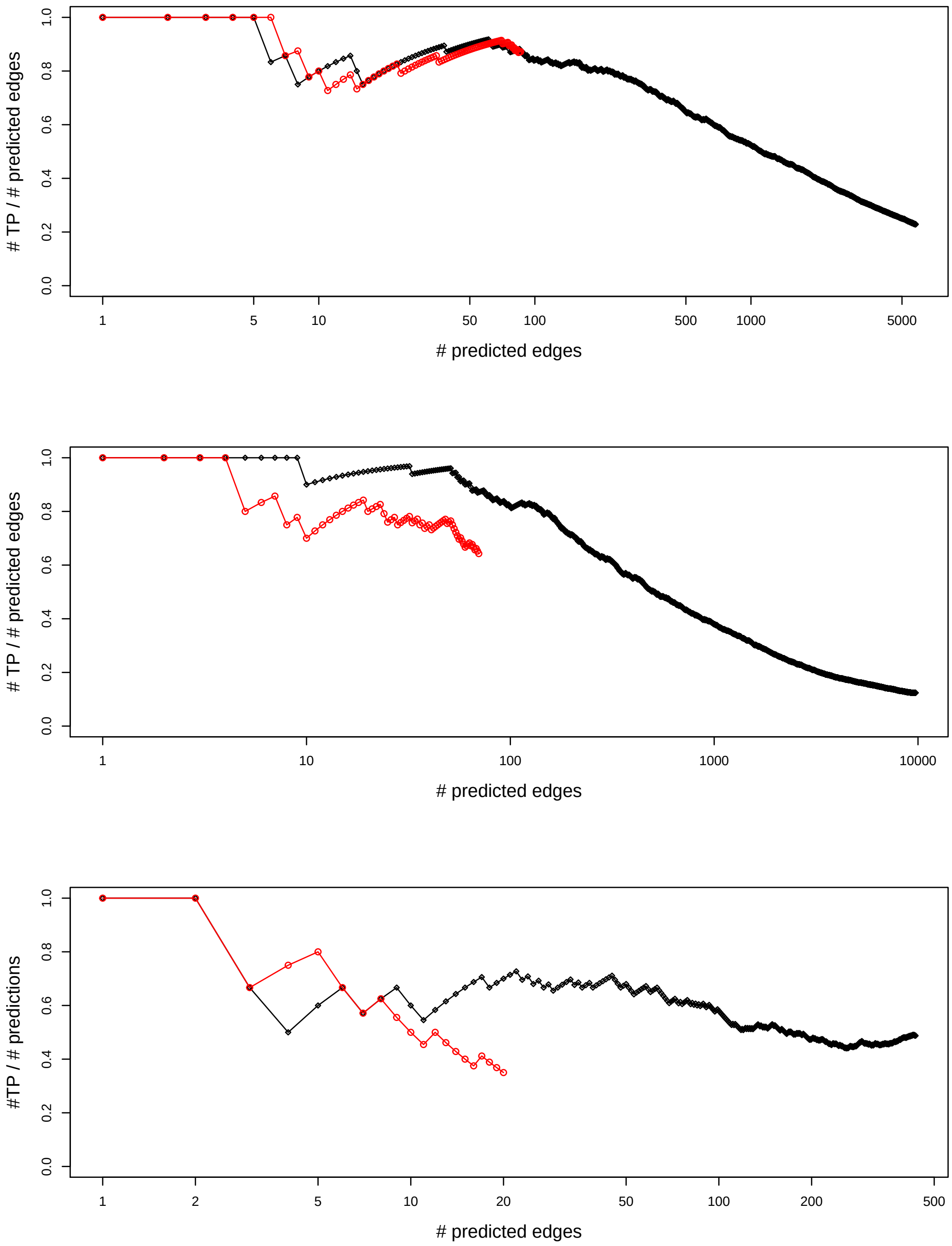

Figure S5: Fraction of true positive (TP) contacts amongst the first predicted pairs ranked by `MIIC` (red curves) and plmDCA (black curves) for three protein structures: *1nxs* (Figs S3 & S4), *1a3a:a* and *1mb6:a*. PlmDCA predictions make use of the full datasets which requires a reweighting scheme to compensate for sampling biases of similar sequences. By contrast, `MIIC` results are based on reduced datasets filtering out sequences with more than 70% identity. This corresponds to reduced datasets including 12,533 out of 63,624 sequences for *1nxs* (upper panel), 2,897 out of 31,922 sequences for *1a3a:a* (middle panel) and 53 out of 246 sequences for *1mb6:a* (lower panel). Note, however, that `MIIC` predicts a finite number of contacts, while plmDCA ranks predictions without a clear cut-off.