BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–3

Supplementary information to the article: Gene network inference by probabilistic scoring of relationships from a factorized model of interactions

Marinka Žitnik[1] and Blaž Zupan[1,2]

[1]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

S1 DERIVATION OF A MAXIMUM A *POSTERIORI* ESTIMATOR FOR A FACTORIZED MODEL

We here show that maximizing the posterior probability of gene latent feature vectors and logistic function parametrization conditioned on the observed phenotypes of double mutants is equivalent to solving an optimization problem with an objective function specified by Eq. (2) in the main manuscript. Reader is referred to the main manuscript for the introduction of the used notation and to Salakhutdinov and Mnih (2008); Park *et al.* (2013) for a background on probabilistic factorized models.

The posterior distribution over the gene latent features $\mathbf{U}$ and $\mathbf{V}$ and the parametrized logistic map $\mathbf{\Psi}$ is given by:

$$
\begin{aligned}
&p(\mathbf{U},\mathbf{V},\mathbf{\Psi}|\mathbf{G},\sigma^2_{\mathbf{G}},\sigma^2_{\mathbf{U}},\sigma^2_{\mathbf{V}},\sigma^2_{\mathbf{\Psi}}) \propto \\
&\propto p(\mathbf{G}|\mathbf{U},\mathbf{V},\mathbf{\Psi},\sigma^2_{\mathbf{G}})p(\mathbf{U}|\sigma^2_{\mathbf{U}})p(\mathbf{V}|\sigma^2_{\mathbf{V}})p(\mathbf{\Psi}|\sigma^2_{\mathbf{\Psi}}) = \\
&= \prod_{u=1}^{n}\prod_{v=1}^{n}(\mathcal{N}(\mathbf{G}_{u,v}|g(\mathbf{U}_u^T\mathbf{V}_v;\mathbf{\Psi}_{u,v}),\sigma^2_{\mathbf{G}}))^{I^{\mathbf{G}}_{u,v}}\times \\
&\times \prod_{u=1}^{n}\mathcal{N}(\mathbf{U}_u|\mathbf{0},\sigma^2_{\mathbf{U}}\mathbf{I}) \times \prod_{v=1}^{n}\mathcal{N}(\mathbf{V}_v|\mathbf{0},\sigma^2_{\mathbf{V}}\mathbf{I})\times \\
&\times \prod_{i=1}^{3}\prod_{u=1}^{n}\prod_{v=1}^{n}(\mathcal{N}(\mathbf{\Psi}^{(i)}_{u,v}|1,\sigma^2_{\mathbf{\Psi}}\mathbf{I}))^{I^{\mathbf{G}}_{u,v}}. \qquad (1)
\end{aligned}
$$

We find a point estimate of unknown $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{\Psi}$ by maximizing the log of the posterior distribution in Eq. (1):

$$
\begin{aligned}
&\ln p(\mathbf{U},\mathbf{V},\mathbf{\Psi}|\mathbf{G},\sigma^2_{\mathbf{G}},\sigma^2_{\mathbf{U}},\sigma^2_{\mathbf{V}},\sigma^2_{\mathbf{\Psi}}) = \\
&= -\frac{1}{2\sigma^2_{\mathbf{G}}}\sum_{u=1}^{n}\sum_{v=1}^{n}I^{\mathbf{G}}_{u,v}(\mathbf{G}_{u,v}-g(\mathbf{U}_u^T\mathbf{V}_v;\mathbf{\Psi}_{u,v}))^2- \\
&-\frac{1}{2\sigma^2_{\mathbf{U}}}\sum_{u=1}^{n}\mathbf{U}_u^T\mathbf{U}_u-\frac{1}{2\sigma^2_{\mathbf{V}}}\sum_{v=1}^{n}\mathbf{V}_v^T\mathbf{V}_v- \\
&-\frac{1}{2\sigma^2_{\mathbf{\Psi}}}\sum_{i=1}^{3}\sum_{u=1}^{n}\sum_{v=1}^{n}I^{\mathbf{G}}_{u,v}(\mathbf{\Psi}^{(i)}_{u,v}-1)^2- \\
&-\frac{1}{2}(\sum_{u=1}^{n}\sum_{v=1}^{n}I^{\mathbf{G}}_{u,v})\ln\sigma^2_{\mathbf{G}}-\frac{1}{2}nk\ln\sigma^2_{\mathbf{U}}-\frac{1}{2}nk\ln\sigma^2_{\mathbf{V}}- \\
&-\frac{1}{2}(\sum_{i=1}^{3}\sum_{u=1}^{n}\sum_{v=1}^{n}I^{\mathbf{G}}_{u,v})\ln\sigma^2_{\mathbf{\Psi}}+\mathcal{C}, \qquad (2)
\end{aligned}
$$

where $\mathcal{C}$ is a constant. If we fix the hyperparameters (i.e. variances $\sigma^2_{\mathbf{G}}$, $\sigma^2_{\mathbf{U}}$, $\sigma^2_{\mathbf{V}}$, $\sigma^2_{\mathbf{\Psi}}$) and set $\lambda_{\mathbf{U}} = \sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{U}}$, $\lambda_{\mathbf{V}} = \sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{V}}$ and $\lambda_{\mathbf{\Psi}} = \sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{\Psi}}$ then maximizing the log posterior is equivalent to minimizing the following objective function, which is the same as Eq. (2) from the main manuscript:

$$
\begin{aligned}
\mathcal{L}(\mathbf{G},\mathbf{U},\mathbf{V},\mathbf{\Psi}) \quad = \quad & \frac{1}{2}\sum_{u=1}^{n}\sum_{v=1}^{n}I^{\mathbf{G}}_{u,v}(\mathbf{G}_{u,v}-g(\mathbf{U}_u^T\mathbf{V}_v;\mathbf{\Psi}_{u,v}))^2+ \\
&+\frac{\lambda_{\mathbf{U}}}{2}\sum_{u=1}^{n}\mathbf{U}_u^T\mathbf{U}_u+\frac{\lambda_{\mathbf{V}}}{2}\sum_{v=1}^{n}\mathbf{V}_v^T\mathbf{V}_v+ \\
&+\frac{\lambda_{\mathbf{\Psi}}}{2}\sum_{i=1}^{3}\sum_{u=1}^{n}\sum_{v=1}^{n}I^{\mathbf{G}}_{u,v}(\mathbf{\Psi}^{(i)}_{u,v}-1)^2.
\end{aligned}
$$

S2 QUANTITATIVE ANALYSIS OF GENE ORDERING

In the article, we assess the accuracy of predicting the order of genes by comparing it to the order in a known pathway. We score the comparison by a standard measure of an AUC, an area under the receiver operating characteristic (ROC). Here, we explicitly show the corresponding ROC curves for both epistasis-analysis approaches considered. Fig. S1 shows the ROC curves for the ordering from KEGG pathways, and Fig. S2 for N-linked glycosylation pathway.

S3 PREDICTION OF ALLEVIATING GENETIC INTERACTIONS

We observe that the probability of alleviation predicted by Réd is correlated to the strength of alleviating effects of a gene pair (Fig. S3).

S4 SENSITIVITY AND REPEATABILITY ANALYSIS

We analyze the sensitivity of Réd to reduced measurement precision by introducing increasing levels of random noise to the data set of Jonikas *et al.* (2009) and, for each noise level, re-running inference by Réd with a fixed initialization of matrix factors. For every measurement of a single and double mutant in the data set we sample the noise component from a Gaussian distribution with zero

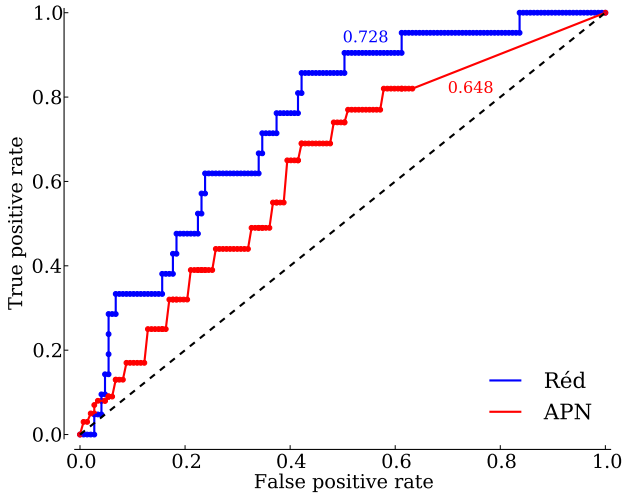

Fig. S1. The ROC curves for the prediction of gene ordering in KEGG pathways by Réd, our proposed approach, and a Bayesian learning method APN (Battle *et al.*, 2010)). Each curve is annotated with its corresponding area under the curve (AUC).

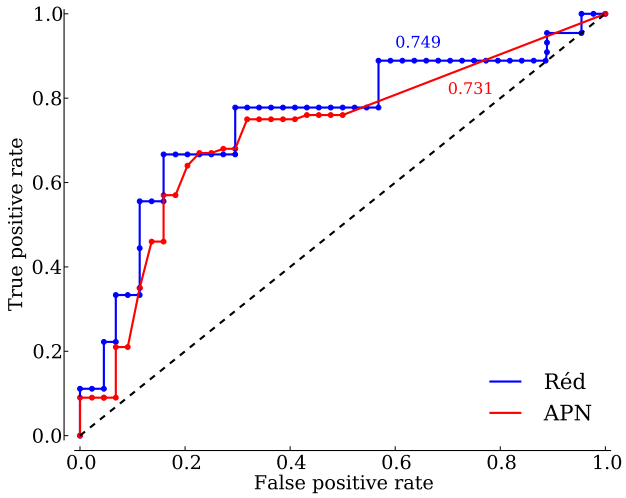

Fig. S2. The ROC curves for the prediction of the edges in the N-linked glycosylation pathway by Réd, our proposed approach, and a Bayesian learning method APN (Battle *et al.*, 2010). Each curve is annotated with its corresponding area under the curve (AUC).

mean and standard deviation s, and add this value to the original measurement. For each run, using a specific value for s, we compare all estimates in $\mathbf{P}$ to its original, noise-free estimates. Fig. S4 shows the correlation between the original estimates and estimates inferred from the noisy data set. The results suggest that good probability estimates of network relationships between genes are possible even in settings with increased noise. Thus, Réd could also infer accurate

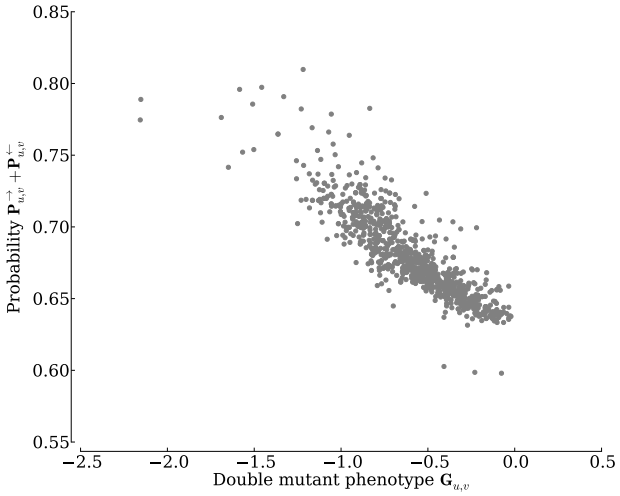

Fig. S3. Probabilities of alleviating gene pairs predicted by Réd are correlated with the strength of alleviating interactions (Spearman $r = -0.704$, p-value $< 1 \times 10^{-100}$). Notice that alleviation corresponds to negative interaction values. Réd assigns higher scores to gene pairs with stronger alleviating effects that to gene pairs that interact moderately.

networks from data that includes more noise than otherwise present in the data set by Jonikas *et al.* (2009).

For twenty runs of Réd learning with different initializations of matrix factors $\mathbf{U}$ and $\mathbf{V}$, we estimate $\mathbf{P}$ for the edges potentially connecting each pair of genes. For every run we compare all probability estimates in $\mathbf{P}$ to the corresponding estimates from every other run. The maximum difference for any two runs and for any pair of genes is less than 1×10^{-8}, demonstrating that Réd estimates are highly repeatable and that the performance of Réd does not substantially vary with initialization of the latent factors.

Similarly, we run Réd several times for different values of the latent dimension k ($k \in \{40, 60, 80, 100, 120\}$). We compare the corresponding probability estimates in $\mathbf{P}$ from every two runs. The mean difference for any two runs and for any edge is less than 1×10^{-3} and the standard deviation is less than 1×10^{-2}. Thus, Réd is robust and performs well on the data by Jonikas *et al.* (2009) for a broad range of sensible values for the latent dimension.

REFERENCES

Battle, A., Jonikas, M. C., Walter, P., Weissman, J. S., and Koller, D. (2010). Automated identification of pathways from quantitative genetic interaction data. *Molecular Systems Biology*, **6**.

Jonikas, M. C., Collins, S. R., Denic, V., Oh, E., Quan, E. M., Schmid, V., Weibezahn, J., Schwappach, B., Walter, P., Weissman, J. S., *et al.* (2009). Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science*, **323**(5922), 1693–1697.

Park, S., Kim, Y.-D., and Choi, S. (2013). Hierarchical bayesian matrix factorization with side information. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1593–1599.

Salakhutdinov, R. and Mnih, A. (2008). Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems 20*, pages 1257–1264.

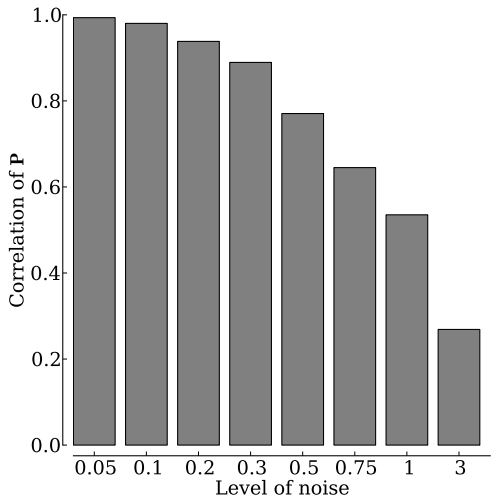

Fig. S4. Sensitivity of $\mathbf{P}$ to measurement noise. We vary the level of Gaussian noise introduced into phenotypic measurements of single and double mutants for the Jonikas *et al.* data and compute the correlation between $\mathbf{P}$ as estimated from the original or noise induced data.