BIOINFORMATICS

Vol. 30 ISMB 2014, pages i246–i254
doi:10.1093/bioinformatics/btu287

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, SI-1000 Ljubljana, Slovenia and [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

ABSTRACT

Motivation: Epistasis analysis is an essential tool of classical genetics for inferring the order of function of genes in a common pathway. Typically, it considers single and double mutant phenotypes and for a pair of genes observes whether a change in the first gene masks the effects of the mutation in the second gene. Despite the recent emergence of biotechnology techniques that can provide gene interaction data on a large, possibly genomic scale, few methods are available for quantitative epistasis analysis and epistasis-based network reconstruction.

Results: We here propose a conceptually new probabilistic approach to gene network inference from quantitative interaction data. The approach is founded on epistasis analysis. Its features are joint treatment of the mutant phenotype data with a factorized model and probabilistic scoring of pairwise gene relationships that are inferred from the latent gene representation. The resulting gene network is assembled from scored pairwise relationships. In an experimental study, we show that the proposed approach can accurately reconstruct several known pathways and that it surpasses the accuracy of current approaches.

Availability and implementation: Source code is available at http://github.com/biolab/red.

Contact: blaz.zupan@fri.uni-lj.si

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Epistasis analysis is a tool of classical genetics for inferring the order of genes in pathways from mutant-based phenotypes (Avery and Wasserman, 1992; Botstein and Maurer, 1982). Epistasis asserts that two genes interact if the mutation in one gene masks the effects of perturbations in the other gene. Then, assuming a common pathway, the first masking gene would be downstream, and the products of the second gene would regulate the expression of the first one (Avery and Wasserman, 1992; Cordell, 2002; Huang and Sternberg, 1995; Roth *et al.*, 2009). Epistasis analysis uncovers the relationship between a pair of genes. Its logic can be further extended to uncover parallelism, where both genes have an effect on the phenotype but where there is no epistasis (Battle *et al.*, 2010; Zupan *et al.*, 2003) (Fig. 1). Uncovered pairwise relationships in a group of genes can give rise to a reconstruction of more complex multi-gene networks. An enlightening demonstration of the power of epistasis for assembly of gene networks is for instance a reconstruction of a four-gene cell death pathway in *Caenorhabditis elegans* (Metzstein *et al.*, 1998).

*To whom correspondence should be addressed.

Emergent technologies from molecular biology that record phenotypes of single and double mutants at a large, possibly genomic scale, prompt for the development of systematic approaches for epistasis analysis and pose the need to devise computational tools that support gene network inference. Approaches of mutagenesis by homologous recombination (Collins *et al.*, 2006; Tong *et al.*, 2004) or RNA interference can yield phenotype observations for thousands or even millions of mutants (Costanzo *et al.*, 2010). Several past studies considered mutant assays with qualitative phenotypes (Zupan

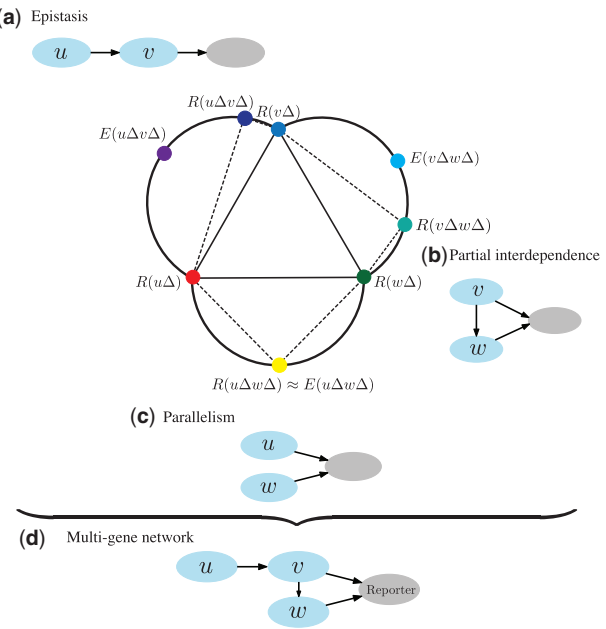


Fig. 1. A hypothetical example of epistasis analysis with three genes, *u*, *v* and *w*. Nodes in the central graph represent mutant phenotypes. The phenotypic difference between a double knockout [e.g. $R(u\Delta v\Delta)$] and a single knockout mutant [e.g. $R(v\Delta)$] is represented with the length of the corresponding dotted edge. Expected double mutant phenotypes, which assume no interaction between genes (see also Section 2.1), are denoted with *E* [e.g. $E(u\Delta v\Delta)$]. A double mutant $u\Delta v\Delta$ (**a**) has a phenotype similar to that of a single mutant $v\Delta$, which indicates that *v* is epistatic to *u*. From the activity of genes *v* and *w* (**b**) we conjecture that gene *v* partially depends on gene *w*, i.e. *v* also acts through a separate pathway because their double mutant $v\Delta w\Delta$ has a phenotype that is equally similar to the single knockout $R(w\Delta)$ and the expected phenotype $E(v\Delta w\Delta)$. The phenotype of double knockout $u\Delta w\Delta$ (**c**) is close to the expected phenotype of $u\Delta w\Delta$, $E(u\Delta w\Delta)$, which may be explained by *u* and *w* acting independently in parallel pathways. Gene ordering from these three relations is preserved in the joint network (**d**), which is a candidate pathway of genes *u*, *v* and *w*

et al., 2003), quantitative fitness scores (Battle *et al.*, 2010; Beerenwinkel *et al.*, 2007; Drees *et al.*, 2005; Phenix *et al.*, 2011, 2013; St Onge *et al.*, 2007) or even whole-genome transcriptional profiles (Hughes, 2005; Van Driessche *et al.*, 2005). Majority of these studies present gene networks as collections of directly observed pairwise interactions (e.g. Phenix *et al.*, 2013; St Onge *et al.*, 2007) and do not propose a generally applicable formalism to model the data. Only few general-purpose algorithms for inference of epistatic networks have been proposed. Zupan *et al.* (2003) introduced formal rules and inference algorithm to infer different types of relationships between genes, but could treat only qualitative phenotypes and could not handle noise. These limitations were elegantly bypassed by a Bayesian approach of Battle *et al.* (2010) that can handle larger data sets with few hundred genes. This algorithm is to our knowledge also the only modern approach to inference of epistasis networks.

Gene epistasis analysis infers interactions that stem directly from mutant phenotypes. Its causative reasoning is different from other network reconstruction tools that observe correlations between gene profiles (e.g. Ahn *et al.*, 2011; Mohammadi *et al.*, 2012) and infer relationships that are circumstantial (Hughes *et al.*, 2000). Despite the growing body of quantitative genetic interaction data and our ability to collect such data, computational approaches and tools to support epistasis are at best scarce (Battle *et al.*, 2010; Jaimovich and Friedman, 2011; Zhang and Zhao, 2013). Devising methods for inference of gene pathways from mutant-based phenotypes and developing related software tools remains a major challenge of computational systems biology.

We here present a new epistasis analysis-inspired computational approach to infer gene networks from a collection of quantitative mutant phenotypes. We refer to our method as Réd (pronounced as *réd*, meaning 'order' in Slovene). Our work was motivated by the Bayesian learning method of Battle *et al.* (2010), henceforth denoted by activity pathway network (APN), that starts from a random network and then iteratively refines it to best match data-inferred relationships. The model refinement in APN is carried out through a succession of local structural changes of the evolving network. This procedure may substantially depend on (arbitrary) initialization of network structure, and hence requires ensembling across many runs of the algorithm to raise accuracy of the final network.

Our approach is conceptually different from APN. We first simultaneously infer a probabilistic model for the entire set of pairwise relationships. Relationship probabilities serve as preferences for different types of pairwise relationships (e.g. epistasis, parallelism and partial interdependence) used in a single-step construction of a gene network. In contrast to APN's local network changes, Réd applies a global procedure to infer the relationships between genes and does not require ensembling. The probabilistic model of Réd uses matrix completion-derived latent data representation to account for noise and sparsity. Inference of factorized model also includes construction of a gene-specific data transformation to account for the differences in single mutant backgrounds, which may affect the phenotype of double mutants. In an experimental study, we show that both components are necessary for inferring gene networks of high accuracy.

2 MATERIALS AND METHODS

Réd, the proposed gene network reconstruction algorithm (Alg. 1), considers quantitative phenotype measurements over a set of single and double mutants, provides preferential order-of-action scores of possible pairwise relationships and assembles them in a joint gene network. The essential steps of the algorithm are overviewed in Figure 2 and are described in detail below.

2.1 Problem definition

In quantitative analysis of genetic interactions we typically observe pairwise interactions between n genes and measure mutant phenotypes, such as the fitness of an organism or expression of a reporter gene (*Reporter*). Measurements over a set of double knockout mutants are given in a sparse matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ and those of single knockout mutants in a vector $\mathbf{S} \in \mathbb{R}^{n}$. In these matrices, $\mathbf{G}_{u,v}$ quantifies a phenotype of double mutant $u\Delta v\Delta$ and $\mathbf{S}_u$ denotes a phenotype of single mutant $u\Delta$. The expected mutant phenotypes, which represent phenotypes of double mutants in the absence of genetic interactions, are given by a matrix $\mathbf{H}$.

We aim to reconstruct a gene network that is consistent with pairwise gene relationships inferred from $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{S}$. Inputs to network reconstruction are preferential scores for all four modeled gene relationships that include epistasis $u \rightarrow v$, epistasis $u \leftarrow v$, parallelism $v||u$ and partial interdependence $v\triangle u$ (Table 1). Réd represents the scores as $\mathbf{P} = (\mathbf{P}^{\rightarrow}, \mathbf{P}^{\leftarrow}, \mathbf{P}^{||}, \mathbf{P}^{\triangle})$ and computes them from the latent gene representation, which is obtained in the inference of a factorized model.

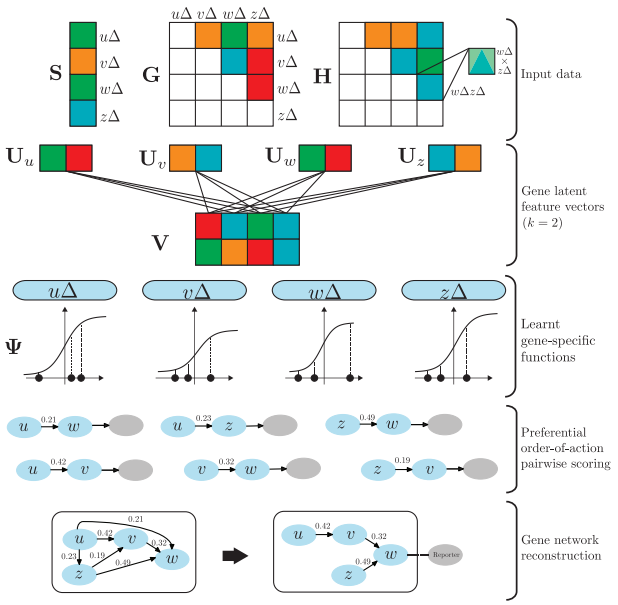
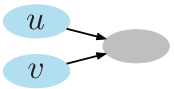
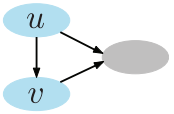


Fig. 2. An overview of Réd, a novel approach for automatic gene network inference from mutant data. Inputs to the preferential order-of-action factorized algorithm of Réd include a matrix of double knockout phenotypes (**G**), a vector of single knockout phenotypes (**S**) and a matrix of expected phenotypes corresponding to the assumption of absent interactions between genes (**H**). Réd estimates a factorized model from **G**, whose gene latent feature vectors capture the global structure of the phenotype landscape, and learns a parametrized logistic map **Ψ**, which is a gene-dependent non-linear mapping from latent to phenotype space. A scoring scheme is then applied to the inferred model to estimate the probabilities of pairwise gene relationships of different types. Finally, a multi-gene network is reconstructed, which aims to minimize the number of violating and redundant edges

Table 1. Probabilistic scoring of gene-gene relationships

Gene–gene relationship	Network structure	Preferential order-of-action score
u and v in a linear pathway, v downstream, gene v is *epistatic* to gene u		$\mathbf{P}^{\rightarrow}_{u,v} = \frac{2}{1+\exp\left(\lvert\hat{\mathbf{G}}_{u,v}-\mathbf{S}_v\rvert\right)}$
u and v in a linear pathway, u downstream, gene u is *epistatic* to gene v		$\mathbf{P}^{\leftarrow}_{u,v} = \frac{2}{1+\exp\left(\lvert\hat{\mathbf{G}}_{u,v}-\mathbf{S}_u\rvert\right)}$
u and v affect the reporter separately		$\mathbf{P}^{\lvert\rvert}_{u,v} = \frac{2}{1+\exp\left(\lvert\hat{\mathbf{G}}_{u,v}-\mathbf{H}_{u,v}\rvert\right)}$
u and v are partially interdependent, each has also a path to the reporter that is independent of the other		$\mathbf{P}^{\triangle}_{u,v} = \frac{2}{1+\exp\left(\lvert\hat{\mathbf{G}}_{u,v}-\frac{1}{2}(\mathbf{H}_{u,v}+\max(\mathbf{S}_u,\mathbf{S}_v))\rvert\right)}$

Given genes u and v, the table shows all four pairwise relationships and their corresponding network structures. These relationships have already been considered by Battle *et al.* (2010) but are here studied with probabilistic scoring functions. See main text for explanation of preferential order-of-action scores.

2.2 Factorized model

To deal with noise and address possibly incomplete input data, Réd estimates probabilities of gene relationships through a factorized model. We use a Bayesian inference approach and formulate the conditional probability of observed double mutant phenotype data, given their latent representation, as follows:

$$p(\mathbf{G}|\mathbf{U},\mathbf{V},\mathbf{\Psi},\sigma^2_{\mathbf{G}}) = \prod_{u=1}^{n}\prod_{v=1}^{n}(\mathcal{N}(\mathbf{G}_{u,v}|g(\mathbf{U}_u^T\mathbf{V}_v;\mathbf{\Psi}_{u,v}),\sigma^2_{\mathbf{G}}))^{I^{\mathbf{G}}_{u,v}},$$

where $\mathcal{N}(x|\mu,\sigma^2)$ is a normal distribution with mean μ and variance σ^2, and $I^{\mathbf{G}}_{u,v}$ indicates whether the phenotypic measurement of $u\Delta v\Delta$ is available.

We assume that the observed phenotype of $u\Delta v\Delta$ is governed by the latent features associated with both genes u and v. To learn the latent features of u and v, we factorize double mutant phenotype data ($\mathbf{G}$) into a product of two low-dimensional latent matrix factors $\mathbf{U}^{k\times n}$ and $\mathbf{V}^{k\times n}$. Their column vectors, $\mathbf{U}_u$ and $\mathbf{V}_v$, represent k-dimensional u-specific and v-specific gene latent feature vectors, respectively. Instead of using linear latent Gaussian model of gene interactions, we pass the dot product $\mathbf{U}_u^T\mathbf{V}_v$ through a parametrized logistic function g. Thus, the model of interaction between genes u and v is represented by the factorized parameter $g(\mathbf{U}_u^T\mathbf{V}_v;\mathbf{\Psi}_{u,v})$. In the factorization, gene interactions depend on each other, as they overlap and share parameters. For instance, given genes u, v and w, their factorized parameters $g(\mathbf{U}_u^T\mathbf{V}_v;\mathbf{\Psi}_{u,v})$ and $g(\mathbf{U}_u^T\mathbf{V}_w;\mathbf{\Psi}_{u,w})$ share a common gene latent feature vector $\mathbf{U}_u$.

Parametrized logistic function g is given by

$$g(x;\psi^{(1)},\psi^{(2)},\psi^{(3)}) = \frac{\psi^{(3)}}{1+\psi^{(1)}\exp\left(-\psi^{(2)}x\right)}$$

and bounds the range of factorized parameters by modeling saturation of the *Reporter*. Here, parameter $\psi^{(3)}$ represents the limiting value of the output past that $g(x;\psi^{(1)},\psi^{(2)},\psi^{(3)})$ cannot grow and $\psi^{(1)}$ represents the number of times that $\mathbf{U}_u^T\mathbf{V}_v$ must grow to reach the value of $\psi^{(3)}$. If $\psi^{(2)}$ is positive, g is increasing in x, otherwise g is a decreasing function. Note that $g(x;1,1,1)$ corresponds to the well-known sigmoid function. For every double mutant $u\Delta v\Delta$, we represent its logistic function parameters in a triple $\mathbf{\Psi}_{u,v} = (\mathbf{\Psi}^{(1)}_{u,v},\mathbf{\Psi}^{(2)}_{u,v},\mathbf{\Psi}^{(3)}_{u,v})$ and define $\mathbf{\Psi}$ to hold the parametrized logistic function representation over all possible double mutants: $\mathbf{\Psi} = (\mathbf{\Psi}^{(1)},\mathbf{\Psi}^{(2)},\mathbf{\Psi}^{(3)})$. We reduce the complexity of this factorized model in Section 2.3 by replacing dense parametrization of $\mathbf{\Psi}$ (one parameter set for every factorized parameter, $|\mathbf{\Psi}| = 3n^2$) with gene-dependent parametrization (one parameter set for every gene, $|\mathbf{\Psi}| = 3n$).

We use a Gaussian prior centered at 1 for logistic function parametrization $\mathbf{\Psi}$ over given phenotypic measurements:

$$p(\mathbf{\Psi}|\sigma^2_{\mathbf{\Psi}}) = \prod_{i=1}^{3}\prod_{u=1}^{n}\prod_{v=1}^{n}(\mathcal{N}(\mathbf{\Psi}^{(i)}_{u,v}|1,\sigma^2_{\mathbf{\Psi}}\mathbf{I}))^{I^{\mathbf{G}}_{u,v}}.$$

For gene latent feature vectors in $\mathbf{U}$ and $\mathbf{V}$ we assume zero-mean Gaussian priors to avoid overfitting:

$$p(\mathbf{U}|\sigma^2_{\mathbf{U}}) = \prod_{u=1}^{n}\mathcal{N}(\mathbf{U}_u|\mathbf{0},\sigma^2_{\mathbf{U}}\mathbf{I}),\ p(\mathbf{V}|\sigma^2_{\mathbf{V}}) = \prod_{v=1}^{n}\mathcal{N}(\mathbf{V}_v|\mathbf{0},\sigma^2_{\mathbf{V}}\mathbf{I}).$$

Through Bayesian inference we derive the posterior probability of gene latent vectors and logistic function parametrization given the available double mutants phenotypes:

$$p(\mathbf{U},\mathbf{V},\mathbf{\Psi}|\mathbf{G},\sigma^2_{\mathbf{G}},\sigma^2_{\mathbf{U}},\sigma^2_{\mathbf{V}},\sigma^2_{\mathbf{\Psi}}) \propto p(\mathbf{G}|\mathbf{U},\mathbf{V},\mathbf{\Psi},\sigma^2_{\mathbf{G}})p(\mathbf{U}|\sigma^2_{\mathbf{U}})\,p(\mathbf{V}|\sigma^2_{\mathbf{V}})p(\mathbf{\Psi}|\sigma^2_{\mathbf{\Psi}}). \quad (1)$$

We select the factorized model according to the maximum *a posteriori* (MAP) estimation by maximizing the log-posterior of Equation (1) over latent feature matrices and logistic function parametrization. The measurement noise variance ($\sigma^2_{\mathbf{G}}$) and prior variances ($\sigma^2_{\mathbf{U}}$, $\sigma^2_{\mathbf{V}}$ and $\sigma^2_{\mathbf{\Psi}}$) are kept fixed. This is equivalent to minimizing the following objective function (see Supplementary Material for a detailed derivation of a MAP estimator), which is a sum of squared errors with quadratic regularization terms:

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

where $\lambda_{\mathbf{U}} = \sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{U}}$, $\lambda_{\mathbf{V}} = \sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{V}}$ and $\lambda_{\mathbf{\Psi}} = \sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{\Psi}}$.

Because $\mathbf{\Psi}$, $\mathbf{U}$ and $\mathbf{V}$ are unknown, the function $\mathcal{L}$ is not convex. In particular, $\mathcal{L}$ is convex in either $\mathbf{U}$ or $\mathbf{V}$ but not in both factors together, which is a known result from matrix factorization studies (Koren *et al.*, 2009; Lee and Seung, 2000). In our study, $\mathcal{L}$ is further coupled by the parametrization of $\mathbf{\Psi}$. Thus, it is unrealistic to expect an algorithm to solve the optimization problem defined by $\mathcal{L}$ in the sense of finding global minimum. We thus estimate latent features and logistic function parameters by finding a local minimum of the objective function $\mathcal{L}$ through

application of gradient descent. Derivatives of $\mathcal{L}$ with respect to gene latent features and logistic parameters are given by the following equations:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{U}_u} = \sum_{v=1}^{n} h(u, v)\mathbf{V}_v g'(\mathbf{U}_u^T\mathbf{V}_v; \boldsymbol{\Psi}_{u,v}) + \lambda_{\mathbf{U}}\mathbf{U}_u, \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}_v} = \sum_{u=1}^{n} h(u, v)\mathbf{U}_u g'(\mathbf{U}_u^T\mathbf{V}_v; \boldsymbol{\Psi}_{u,v}) + \lambda_{\mathbf{V}}\mathbf{V}_v, \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Psi}^{(1)}_{u,v}} = -\frac{h(u, v)\boldsymbol{\Psi}^{(3)}_{u,v}\exp\left(\boldsymbol{\Psi}^{(2)}_{u,v}\mathbf{U}_u^T\mathbf{V}_v\right)}{\left(\exp\left(\boldsymbol{\Psi}^{(2)}_{u,v}\mathbf{U}_u^T\mathbf{V}_v\right) + \boldsymbol{\Psi}^{(1)}_{u,v}\right)^2} + t(u, v, 1), \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Psi}^{(2)}_{u,v}} = \frac{h(u, v)\boldsymbol{\Psi}^{(1)}_{u,v}\boldsymbol{\Psi}^{(3)}_{u,v}\mathbf{U}_u^T\mathbf{V}_v\exp\left(\boldsymbol{\Psi}^{(2)}_{u,v}\mathbf{U}_u^T\mathbf{V}_v\right)}{\left(\exp\left(\boldsymbol{\Psi}^{(2)}_{u,v}\mathbf{U}_u^T\mathbf{V}_v\right) + \boldsymbol{\Psi}^{(1)}_{u,v}\right)^2} + t(u, v, 2), \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Psi}^{(3)}_{u,v}} = \frac{h(u, v)}{1 + \boldsymbol{\Psi}^{(1)}_{u,v}\exp\left(-\boldsymbol{\Psi}^{(2)}_{u,v}\mathbf{U}_u^T\mathbf{V}_v\right)} + t(u, v, 3), \tag{7}$$

where for convenience of notation, $h(u, v)$ is substituted for $h(u, v) = I^{\mathbf{G}}_{u,v}(g(\mathbf{U}_u^T\mathbf{V}_v; \boldsymbol{\Psi}_{u,v}) - \mathbf{G}_{u,v})$, penalty term $t(u, v, i)$ stands for $t(u, v, i) = \lambda_{\boldsymbol{\Psi}} I^{\mathbf{G}}_{u,v}(\boldsymbol{\Psi}^{(i)}_{u,v} - 1)$ and $g'(x; \boldsymbol{\Psi}_{u,v})$ is logistic function derivative with respect to x. Efficiency in training Réd model comes from finding point estimates of model unknowns instead of inferring the full posterior distribution over them.

2.3 Gene-dependent weighting

We further reduce complexity of the model described in the previous section by combining evidence from multiple phenotypic measurements through their latent representation. We replace entrywise (double-mutant-phenotype-dependent) logistic function parametrization $\boldsymbol{\Psi}$ with gene-dependent parametrization that is given by $\boldsymbol{\Psi}^{(i)}_{u,v} \leftarrow \frac{1}{n-1}\sum_w \boldsymbol{\Psi}^{(i)}_{u,w}$ for $i = 1, 2, 3$. This reduces the number of parameters in $\boldsymbol{\Psi}$ that have to be learned from $3n^2$ to $3n$. Intuitively, measurements that involve gene u are not independent from each other but are rather governed by the gene pathways in which u participates. Gene-dependent parametrization of $\boldsymbol{\Psi}$ represents a method of regularization allowing us to remove penalty terms in Equations (5)–(7).

Derivatives of $\boldsymbol{\Psi}$ use only available phenotypic measurements owing to the application of an indicator function [cf. Equations (5)–(7)]. We relax this limitation by considering current estimates of $\mathbf{G}$ when computing the derivatives of $\boldsymbol{\Psi}$. These estimates are given by $\hat{\mathbf{G}}_{u,v} = g(\mathbf{U}_u^T\mathbf{V}_v; \boldsymbol{\Psi}_{u,v})$, where $\mathbf{U}$ and $\mathbf{V}$ are latent matrix factors from the previous iteration of gradient descent (Step 3c in Alg. 1).

2.4 Preferential order-of-action scoring of gene pairs

Probabilities of gene–gene relationships in $\mathbf{P}$ are computed from the inferred phenotypes given by $\hat{\mathbf{G}} = g(\mathbf{U}^T\mathbf{V}; \boldsymbol{\Psi})$, with the rules outlined in Table 1. Estimated probabilities in $\mathbf{P}$ approach 1 when inferred phenotypic values in $\hat{\mathbf{G}}$ are close to the phenotypes, which would be expected if a certain network structure ($\rightarrow$, $\leftarrow$, $||$, $\triangle$) existed between genes, and they slowly vanish when the inferred values deviate from the values expected by a certain type of relationship.

For instance, an epistatic genetic interaction $u \leftarrow v$ is inferred when the trait $\hat{\mathbf{G}}_{u,v}$ of the double mutant $u\Delta v\Delta$ is similar to the single mutant $u\Delta$ phenotype $\mathbf{S}_u$ and the two single mutant phenotypes are different ($\mathbf{S}_u \not\approx \mathbf{S}_v$). This brings $|\hat{\mathbf{G}}_{u,v} - \mathbf{S}_u|$ close to 0 and, consequently, $\mathbf{P}^{\leftarrow}_{u,v}$ close to 1. With different single mutant phenotypes, the expected phenotype $\mathbf{H}_{u,v}$ of the double mutant that assumes no genetic interaction is different from both single mutant phenotypes ($\mathbf{S}_u \not\approx \mathbf{S}_v \Rightarrow \mathbf{S}_v \not\approx \mathbf{H}_{u,v} \wedge \mathbf{S}_u \not\approx \mathbf{H}_{u,v}$), bringing $\mathbf{P}^{||}_{u,v}$ and $\mathbf{P}^{\triangle}_{u,v}$ close to 0. Likewise, the phenotype of $v\Delta$ would be different from the phenotype of the double mutant, bringing $\mathbf{P}^{\rightarrow}_{u,v}$ close to 0.

Cases with less pronounced differences between phenotypes would lead to smaller differences in relationship probabilities. Preferential order-of-action scores generalize the epistasis analysis framework by Avery and Wasserman (1992), wherein the signal and the genes under study were strictly on or off with no intermediate levels of activity. An appealing feature of scores in $\mathbf{P}$ is that they have a direct probabilistic interpretation.

2.5 Multi-gene network inference

Given probabilistic scores of gene–gene network structures in $\mathbf{P}$ from Section 2.4, we reconstruct a detailed multi-gene network that is consistent with the inferred relationship probabilities and contains a minimum number of *violating* and *redundant* edges. Examples of inferred networks are given in Figures 4–7. A network is a weighted directed graph with genes as vertices and directed edges that determine the order of action. A designated vertex represents the observed quantitative trait. A directed edge from u to v is *violating* (Fig. 3a) if there is evidence in $\mathbf{P}$ for both $u \rightarrow v$ and $u \leftarrow v$ (e.g. $\mathbf{P}^{\rightarrow}_{u,v} \approx \mathbf{P}^{\leftarrow}_{u,v}$). A directed edge from u to v is *redundant* (Fig. 3b) if there is evidence in $\mathbf{P}$ that some intermediate gene exists between u and v. That is, u and v are not adjacent in a genetic network but rather u indirectly affects v, i.e. $\mathbf{P}^{\rightarrow}_{u,v}$ captures the extent to which strict weak ordering of u and v holds.

Network inference procedure assigns a level to every gene in a manner that if there is strong evidence in $\mathbf{P}$ that gene u is placed upstream of gene v, that is, if v is epistatic to u, then level(u) > level(v). In the case of stronger evidence of parallelism or partial interdependence between u and v the level(u) $\approx$ level(v). Several genes can be assigned the same level, but a designated vertex corresponding to a phenotype of interest is the only vertex placed on the lowest level.

Inference of a genetic network involves two phases. In the first phase we perform an approximate topological sort through construction of a directed weighted graph. Given genes u and v and the inferred epistasis relationships between them, the direction and weight of a between-level edge are determined by the maximum of the values $\mathbf{P}^{\rightarrow}_{u,v}$ (edge $u \rightarrow v$) and $\mathbf{P}^{\leftarrow}_{u,v}$ (edge $u \leftarrow v$). Given a parallelism or partial interdependence relationship between u and v, a within-level edge is determined by the maximum of the values $\mathbf{P}^{||}_{u,v}$ (no edge between u and v) and $\mathbf{P}^{\triangle}_{u,v}$ (edge $u \rightarrow v$). This graph may contain directed cycles, and finding an exact topological ordering of its vertices with the minimal set of violating edges is a known NP-hard problem (Charbit *et al.*, 2007; Eades *et al.*, 1993). Thus, we proceed in the following way. We select a vertex with no incoming between-

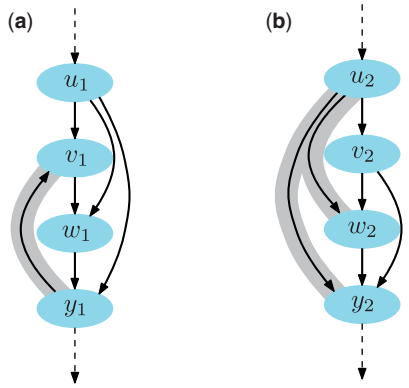


Fig. 3. Illustration of violating (**a**) and redundant (**b**) edges (in gray) in a pathway with four genes. Edge $y_1 \rightarrow v_1$ is violating because there is evidence that v_1 is placed upstream of y_1 ($v_1 \rightarrow w_1$ and $w_1 \rightarrow y_1$) but also that y_1 is upstream of v_1 ($y_1 \rightarrow v_1$). Edge $u_2 \rightarrow w_2$ is redundant because there is evidence of an intermediate gene v_2. Similarly, edge $u_2 \rightarrow y_2$ is redundant because of two intervening genes, v_2 and w_2

level edges, assign that vertex to the currently top-most level and recurse on the graph with that vertex removed. We also look for vertices with no outgoing between-level edges and assign them to the currently lowest level. If in some step multiple vertices have no incoming or outgoing between-level edges, they are assigned the same level. It can happen that all vertices have incoming and outgoing between-level edges. In this case, we select the vertex with the highest differential between weighted incoming between-level degree and weighted outgoing between-level degree.

Alg. 1: Réd, the proposed approach for gene network inference by scoring relationships from a factorized model of interactions.

Input:

- sparse matrix of double mutant phenotypes $\mathbf{G} \in \mathbb{R}^{n \times n}$,
- typical interaction values $\mathbf{H} \in \mathbb{R}^{n \times n}$,
- measured phenotypes of single mutants $\mathbf{S} \in \mathbb{R}^{n}$,
- parameters $\lambda_{\mathbf{U}}$, $\lambda_{\mathbf{V}}$, rates α and β, and rank k.

Output:

- preferential order-of-action score matrices $\mathbf{P}$,
- completed matrix $\hat{\mathbf{G}}$,
- gene-dependent logistic function parametrization $\boldsymbol{\Psi}$,
- inferred gene network for a gene subset of interest.

1. Initialize $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})^{k \times n}$ and $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})^{k \times n}$.
2. Initialize $\boldsymbol{\Psi}^{(i)}$ as $\mathbf{1}_{n \times n}$ for $i = 1, 2, 3$.
3. Repeat until convergence:
   a. Compute $\frac{\partial \mathcal{L}}{\partial \mathbf{U}}$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{V}}$ with Equation (3) and (4), respectively.
   b. Update $\mathbf{U} \leftarrow \mathbf{U} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{U}}$ and $\mathbf{V} \leftarrow \mathbf{V} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{V}}$.
   c. Compute $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Psi}^{(i)}}$ for $i = 1, 2, 3$ using Equations (5)–(7), respectively. Substitute $h(u, v)$ therein with $h(u, v) = g(\mathbf{U}_u^T \mathbf{V}_v; \boldsymbol{\Psi}_{u,v}) - \mathbf{X}_{u,v}$, where $\mathbf{X}_{u,v} = \mathbf{G}_{u,v}$ if $I^{\mathbf{G}}_{u,v} = 1$ and $\mathbf{X}_{u,v} = \hat{\mathbf{G}}_{u,v}$ otherwise. Here, $\hat{\mathbf{G}}_{u,v}$ is computed using the latent matrix factors from the previous iteration.
   d. Update $\boldsymbol{\Psi}^{(i)} \leftarrow \boldsymbol{\Psi}^{(i)} - \beta \frac{\partial \mathcal{L}}{\partial \boldsymbol{\Psi}^{(i)}}$ for $i = 1, 2, 3$.
   e. Set gene-dependent weights $\boldsymbol{\Psi}^{(i)}_{u,v} \leftarrow \frac{1}{n-1} \sum_{w} \boldsymbol{\Psi}^{(i)}_{u,w}$ for $i = 1, 2, 3$ and $\forall u, v$.
4. Compute preferential order-of-action scores $\mathbf{P}^{i}_{u,v}$ for $i \in \{\rightarrow, \leftarrow, ||, \triangle\}$ and $\forall u, v$ using Equations from Table 1.
5. Normalize $\mathbf{P}^{i}_{u,v} \leftarrow \mathbf{P}^{i}_{u,v} / \sum \mathbf{P}^{j}_{u,v}$ for $i \in \{\rightarrow, \leftarrow, ||, \triangle\}$ and $\forall u, v$.
6. Compute $\hat{\mathbf{G}}_{u,v} = g(\mathbf{U}_u^T \mathbf{V}_v; \hat{\boldsymbol{\Psi}}_{u,v})$.
7. Given a gene subset of interest, infer a network (Section 2.5).

In the second phase of gene network inference we retain within-level edges and those edges that link adjacent levels and are directed downward. The latter procedure eliminates violating edges. As a final step, we remove redundant edges according to their definition above.

3 DATA AND EXPERIMENTAL SETUP

We assess the accuracy of Réd by applying our inference approach to the datasets of Jonikas *et al.* (2009) and Surma *et al.* (2013) and compare results to known or partially known networks. Experiments that use data from Jonikas *et al.* closely follow the setup by Battle *et al.* and use the same datasets and reference pathways.

3.1 Mutant phenotype data

Jonikas *et al.* (2009) measured unfolded protein response (UPR) levels in single and double mutants to systematically characterize functional interdependence of yeast genes with roles in endoplasmic reticulum (ER) folding. The dataset contains 444 genes that caused high UPR reporter inductions. The interaction data include phenotypes of 42 240 distinct double mutants (matrix **G**) corresponding to 43% of all possible double mutants. Jonikas *et al.* also computed typical (i.e. expected) values of genetic interactions for every double mutant (matrix **H**). They considered multiplicative neutrality function (Mani *et al.*, 2008) and computed it using reporter levels of pairs of single mutants, modified by a Hill function to account for the saturation of the reporter signal.

Surma *et al.* (2013) considered 741 genes and observed the growth phenotype (colony size) for all pairs of double mutants. In total, after filtering out unreliable measurements, their dataset comprises 251 383 double mutant fitness scores. We computed single mutant scores by averaging across all scores of double mutants that included mutations of the corresponding genes. We considered multiplicative model to calculate the expected fitness of a double mutant in the absence of a genetic interaction.

3.2 Gene pathways

We compare gene networks inferred by Réd to a number of known or partially known cellular pathways that include genes whose perturbations are measured by Jonikas *et al.*:

- The N-linked glycosylation pathway consisting of 10 genes whose true ordering is known (Helenius and Aebi, 2004),
- The ER-associated degradation (ERAD) pathway for which many functional interdependencies between its member genes are known,
- Tail-anchored (TA) protein biogenesis machinery consisting of TA proteins important for transmembrane trafficking and the recently discovered GET pathway (Bozkurt *et al.*, 2009; Schuldiner *et al.*, 2008; Stefanovic and Hegde, 2007).

We also compare Réd's networks to well-characterized cellular pathways of phospholipid biosynthesis whose gene mutants are measured by Surma *et al.* and that include the following:

- The Kennedy pathway involved in the synthesis of phosphatidylethanolamine and phosphatidylcholine (PC), and
- The phosphatidylserine to PC conversion pathway.

3.3 Experimental setup

In the first part of the experiments, we use mutant phenotype data to qualitatively evaluate the reconstruction of five gene pathways from Section 3.2. In the second part of the experiments, we evaluate the accuracy of gene ordering through three different setups. In the first two setups, the data-inferred gene ordering was compared with the known pathways. In the third setup, we use cross-validation to estimate the accuracy of

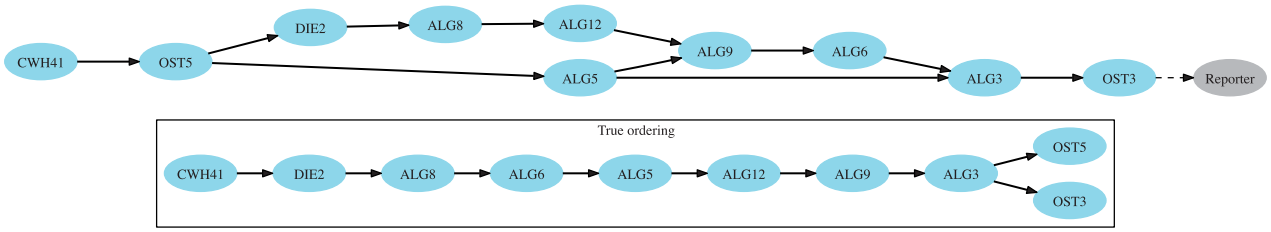


Fig. 4. Gene network of the N-linked glycosylation pathway inferred by Réd. For reference, we show the true ordering of this pathway (Helenius and Aebi, 2004) as adapted from Battle *et al.* (2010). The inferred gene network reflects many correct gene placements

prediction of gene interaction scores with the following experiments:

1. Battle *et al.* (2010) provided 168 test gene pairs (v, u) from common KEGG pathways (Kanehisa *et al.*, 2008). For 21 gene pairs v is known to be upstream of u, and for 147 gene pairs v is not known to be upstream of u. Given a gene pair, Réd predicted the probability of epistasis as $\mathbf{P}^{\rightarrow}_{u,v}/(\mathbf{P}^{\rightarrow}_{u,v}+\mathbf{P}^{\leftarrow}_{u,v})$, and the accuracy of predictions on entire set of 168 gene pairs.
2. Using the setup from Battle *et al.* we evaluate the accuracy of prediction of direct edges $u \rightarrow v$ in the N-linked glycosylation pathway (Fig. 4) based on the model-estimated probability of epistasis $\mathbf{P}^{\rightarrow}_{u,v}$.
3. We estimate the accuracy when predicting that two genes are in epistasis, that is, $u \rightarrow v$ or $v \rightarrow u$. Note that in the literature this relationship is also referred to as an *alleviating interaction*, where the phenotype of a double mutant is less severe than expected from the phenotypes of the corresponding single mutants (Jonikas *et al.*, 2009; Mani *et al.*, 2008). For the data from Jonikas *et al.* this means that the double-mutant cell responds to ER stress surprisingly better than how the ER stress would typically be mitigated. The data for this experiment were preprocessed according to the procedure described by Battle *et al.* A positive set included gene pairs (u, v) with significant alleviating genetic interactions, for which the observed phenotype (interaction score) was negative with a magnitude greater than $|\mathbf{G}_{u,v} - \max(\mathbf{S}_u, \mathbf{S}_v))|$ (see St Onge *et al.*, 2007). It was further required that the double-mutant phenotype data contained a sufficient number of observations that included $u\Delta$ or $v\Delta$, such that the geometric mean of such measurements for u and for v was at least 180. There are 2723 gene pairs in the data of Jonikas *et al.* that match these criteria. In each test run, we form a test set with a random selection of 5% of the positive gene pairs and a negative set of equal size of gene pairs that fail to satisfy the selection criteria. We remove the test data from the interaction score matrix $\mathbf{G}$, and predict whether a test gene pair is alleviating using the probability that u and v occur together in a linear pathway, i.e. $\mathbf{P}^{\rightarrow}_{u,v}+\mathbf{P}^{\leftarrow}_{u,v}$. We report an averaged accuracy across 10 different test runs.

We characterize the accuracy of predictions through the area under the receiver-operating characteristic curve (AUC), with a baseline of 0.5 (random networks) and a perfect score of 1.0 (inferred networks that are identical to gold standard—known networks).

We compare Réd, our network inference approach, with a recently published Bayesian approach by Battle *et al.* They developed preference scoring functions over all possible pairwise gene relationships and applied annealed importance sampling to reconstruct high scoring multi-gene networks. Their method (referred here as APN) was shown to be superior to a number of other approaches that can infer networks from gene interaction data by Jonikas *et al.* These other approaches include baseline techniques such as Pearson correlation of genetic interaction profiles and raw interaction values as well as more sophisticated techniques such as Gaussian process regression (GP; Williams and Rasmussen, 1996), a method that uses the correlation of observed interaction profiles, the diffusion kernel method (DK; Qi *et al.*, 2008) and GenePath (Zupan *et al.*, 2003). For brevity, we therefore focus on comparing our method with APN, which was run with default parameters as chosen by Battle *et al.* for the dataset of Jonikas *et al.*, but we also report the accuracies achieved by GP and DK.

Two essential components of Réd are latent representation of gene interactions and their transformation through the logistic function. To test the extent to which the performance of Réd depends on these two components we also run experiments where the algorithm infers probabilities and makes predictions from raw (not factorized) phenotypes, and where the latent representation is used without logistic transformation. We refer to these two approaches as RAW and MF, respectively.

In all experiments with data from Jonikas *et al.*, the parameters of Réd are set as $\lambda_U = \lambda_V = 1 \times 10^{-4}$, $\beta = 0.1$, $\alpha = 0.1$, $k = 100$. The same parameters are used on data from Surma *et al.* with the exception of $\alpha = 1 \times 10^{-3}$ and $k = 50$, which were selected to minimize the normalized root mean square error of $\hat{\mathbf{G}}$. This choice of regularization parameters and learning rates is common (cf. Min and Lee, 2005; Pedregosa *et al.*, 2011). We also show (see Supplementary Material; Supplementary Fig. S4) that the performance of Réd does not critically depend on the rank of factorization k. Réd's optimization by gradient descent is terminated when the Frobenius distance between $\mathbf{G}$ and $\hat{\mathbf{G}}$ over known values fails to decrease between the two consecutive iterations of optimization.

4 RESULTS AND DISCUSSION

4.1 Reconstruction of a known gene pathway from data by Jonikas *et al.* (2009)

We analyzed the ability of Réd to reconstruct the known N-linked glycosylation pathway. Figure 4 shows the inferred network next to the known pathway as reported by Helenius and

Aebi (2004). Genes *CWH41, DIE2* and *ALG8* are correctly placed such that they are dependent on the other genes. Also, *ALG12* is placed upstream of *ALG9*, which is also upstream of *ALG3*. *OST3* is correctly placed downstream, but *OST5* is incorrectly placed, likely because double-mutant data with the other ALG genes were not available. Surprisingly, Réd correctly placed *CWH41*, a gene that encodes glucosidase I, an integral membrane protein of the ER involved in sensing ER stress (Romero *et al.*, 1997), at the beginning of the pathway despite mild downstream effects observed in *CWH41* mutants. Note that the interaction profile of *CWH41* is only moderately correlated with those of ALG genes, and thus, *CWH41* was not clustered together with them (Jonikas *et al.*, 2009). We hence conclude that Réd inference of the N-linked glycans synthesis pathway was successful with a network that closely resembles that reported in the literature.

4.2 Reconstruction of known gene pathways from data by Surma *et al.* (2013)

We applied Réd to mutant data by Surma *et al.* to reconstruct two thoroughly studied pathways of phospholipid biosynthesis. Réd's ordering of genes in the phosphatidylserine to PC conversion pathway is fully consistent with the reference pathway (Fig. 5a). In the Kennedy pathway, Réd correctly placed *PCT1* upstream of *CPT1* and *CKI1* upstream of *CPT1* with high confidence (Fig. 5b), but it misplaced gene pair *PCT1* and *CKI1* likely owing to the ambiguity in the data. However, as Réd performs global reasoning by combining evidence from all measurements, it handled the data uncertainty by assigning PCT1 → CKI1 structure the lowest score in the reconstruction of the Kennedy pathway.

4.3 Reconstruction of partially known gene pathways

Jonikas *et al.* (2009) identified several pathways that are important for ER protein folding. Of these, the pathways for ERAD and TA protein insertion were considered in Battle *et al.* (2010). Réd-inferred networks for these two pathways are shown in Figures 6 and 7. The solid edges in these figures are those inferred by our algorithm, while the dotted edges indicate gene interactions reported in the literature (Battle *et al.*, 2010; Carvalho *et al.*, 2006; Clerc *et al.*, 2009; Jonikas *et al.*, 2009; Kim *et al.*, 2005; Nakatsukasa and Brodsky, 2008).

The ordering of inferred networks is entirely consistent with the partially known gene pathways. For instance, in the network for the ERAD pathway (Fig. 6), the upstream placement of *MNL1* to *YOS9* is consistent with existing data showing that *MNL1* generates the sugar species recognized by *YOS9* (Clerc *et al.*, 2009). Also, *MNL1*, *YOS9*, *DER1* and *USA1* are placed upstream of *HRD3* and *HRD1*, which is compatible with data showing that degradation of certain substrates requires all six components (Carvalho *et al.*, 2006; Kim *et al.*, 2005; Nakatsukasa and Brodsky, 2008). For the TA protein insertion pathway, Réd inferred a network (Fig. 7) that placed the poorly characterized protein *SGT2* upstream of the TA protein biogenesis machinery components according to its function in the insertion of TA proteins into membranes (Battle *et al.*, 2010).

Similarly, positive results of network inference are also reported in (Battle *et al.*, 2010). Their method inferred a number

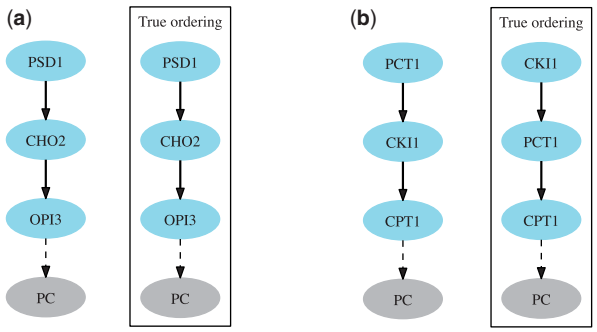


Fig. 5. Gene networks of the phosphatidylserine to PC conversion pathway (**a**) and the Kennedy pathway (**b**) as inferred by Réd. For reference, we show the true orderings in both pathways adapted from Surma *et al.* (2013). Réd correctly and with high confidence ($\mathbf{P} > 0.80$) inferred all three pairwise gene relationships of the PC conversion pathway. It also correctly predicted two out of three gene relationships of the Kennedy pathway with the wrong prediction (PCT1 → CKI1) being assigned a low confidence ($\mathbf{P} = 0.25$)

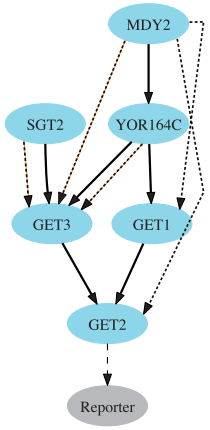


Fig. 7. Gene network inferred by Réd that represents the likely ordering of genes belonging to the TA protein biogenesis machinery (solid edges). Known relationships between genes are denoted by dotted edges. Note that the predicted ordering strongly reflects known interdependencies between genes

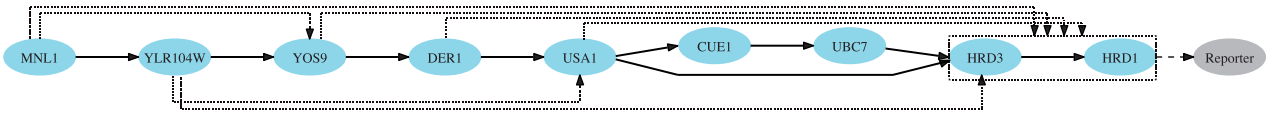


Fig. 6. The ERAD pathway predicted by Réd is shown by solid edges. Placement of genes in the inferred network is consistent with known interdependencies (dotted edges)

Table 2. The predictive accuracy (AUC) of gene ordering by a Bayesian learning method [APN; Battle *et al.* (2010)], Réd, our proposed approach, and its simplified variants: without factorization (RAW) and with factorization but in the absence of transformation by logistic function (MF)

Prediction	AUC			
	RAW	MF	APN	Réd
KEGG pathway ordering	0.563	0.583	0.648	0.728
N-linked glycosylation pathway	0.591	0.638	0.731	0.749

Table 3. Prediction of unknown alleviating genetic interactions

Prediction	AUC				
	MF	DK	APN	GP	Réd
Alleviating genetic interactions	0.723	0.759	0.783	0.862	0.906

We report the accuracy of predicted interactions based on the DK (Qi *et al.*), predictions based on latent representation obtained with standard two-factor matrix factorization (MF), APNs learned through a Bayesian method by Battle *et al.*, predicted genetic interaction values from GP (Williams and Rasmussen) that uses the correlation of observed interaction profiles, and Réd, our proposed approach.

of candidate networks of which the best-scored were shown to be partially consistent with known gene interdependencies. In contrast, for each pathway, Réd inferred a single network that is entirely consistent with known gene relationships.

4.4 Quantitative analysis of gene ordering

Table 2 reports the accuracies of gene ordering prediction obtained by four different algorithms, Réd, APN and two simplified variants of Réd. In comparison with APN, Réd performs substantially better in predicting the edges of the KEGG pathways and slightly better in predicting the edges of the N-linked glycosylation pathway (Supplementary Figs S1 and S2).

The poor performance of the simplified variants of Réd (RAW and MF) indicates that Réd's latent representation inferred from the factorized model, the non-linear logistic map and gene-dependent weighting are the essential components of Réd. Without any of these, Réd would not be able to achieve the resulting accuracy.

4.5 Prediction of alleviating genetic interactions

Given the training and separate test datasets, we predict whether an interaction is alleviating (see Section 3.3). Table 3 shows that Réd performs substantially better than APN ($P<0.001$). Réd also outperforms standard two-factor matrix factorization (MF) by a large margin, which is an indicator that transformation via a logistic map is essential to the performance of our algorithm. We compare these results with those obtained by GP (Williams and Rasmussen, 1996) using squared exponential autocorrelation model constructed from the genetic interaction profiles, and with the interactions predicted with the DK (Qi *et al.*, 2008). Réd achieves significantly higher accuracy than GP ($P<0.01$) and DK ($P<0.001$), although the difference with GP is small and may be worthy of further study. Note that RAW, a Réd variant without factorization, is not applicable for this experiment, as it does not generalize across gene interaction scores.

We have observed that the probabilities of alleviating gene pairs predicted by Réd are well correlated to the strength of alleviating interactions (Spearman $r = -0.704$, $P < 1 \times 10^{-100}$; Supplementary Fig. S3). Réd scores gene pairs with stronger alleviating effects (negative interaction values with greater magnitude) higher than those that interact moderately.

5 CONCLUSION

Réd is a conceptually new approach for inference of gene networks from quantitative genetic interaction data. It implements a probabilistic epistasis analysis and assembles pairwise relationships into gene networks. In our experiments, Réd was able to reconstruct several known and partially known pathways with accuracy above that of the state-of-the-art approaches. Réd outperforms APN, the state-of-the-art method by Battle *et al.* (2010), both in accuracy and speed, with CPU runtime of only a few minutes compared with APN's 30 min for an inference of a single full network in an ensemble of 500 networks. We also show that Réd's power of generalization comes from its two key components, a factorized model with latent representation of gene interactions and a gene-dependent logistic map of interaction scores.

Our evaluation in this article was computational and thus limited to datasets for which several gene pathways or at least partial gene orderings were available (Battle *et al.*, 2010; Jonikas *et al.*, 2009). Réd can efficiently handle similar datasets as well as much larger ones, such as that from the recent yeast experiments by Costanzo *et al.* (2010). These are also the datasets for which we foresee future applications of Réd and which will require subsequent verification of inferred networks in the wet lab.

ACKNOWLEDGEMENTS

We would like to thank Uroš Petrovič for recommending the data set by Surma *et al.* (2013).

Funding: This work was supported by the grants from the Slovenian Research Agency (P2-0209, J2-5480), EU FP7 (Health-F5-2010-242038), NIH (P01-HD39691) and the Fulbright Scholarship (BZ).

Conflict of Interest: none declared.

REFERENCES

Ahn,J. *et al.* (2011) Integrative gene network construction for predicting a set of complementary prostate cancer genes. *Bioinformatics*, **27**, 1846–1853.

Avery,L. and Wasserman,S. (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet.*, **8**, 312–316.

Battle,A. *et al.* (2010) Automated identification of pathways from quantitative genetic interaction data. *Mol. Sys. Biol.*, **6**, 379.

Beerenwinkel,N. *et al.* (2007) Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC Evol. Biol.*, **7**, 6.

Botstein,D. and Maurer,R. (1982) Genetic approaches to the analysis of microbial development. *Annu. Rev. Genet.*, **16**, 61–83.

Bozkurt,G. *et al.* (2009) Structural insights into tail-anchored protein binding and membrane insertion by Get3. *Proc. Natl Acad. Sci. USA*, **106**, 21131–21136.
Carvalho,P. *et al.* (2006) Distinct ubiquitin-ligase complexes define convergent pathways for the degradation of ER proteins. *Cell*, **126**, 361–373.
Charbit,P. *et al.* (2007) The minimum feedback arc set problem is NP-hard for tournaments. *Comb., Probab. Comput.*, **16**, 1–4.
Clerc,S. *et al.* (2009) Htm1 protein generates the N-glycan signal for glycoprotein degradation in the endoplasmic reticulum. *J. Cell Biol.*, **184**, 159–172.
Collins,S.R. *et al.* (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol.*, **7**, R63.
Cordell,H.J. *et al.* (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, **11**, 2463–2468.
Costanzo,M. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
Drees,B.L. *et al.* (2005) Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol.*, **6**, R38.
Eades,P. *et al.* (1993) A fast and effective heuristic for the feedback arc set problem. *Inf. Process. Lett.*, **47**, 319–323.
Helenius,A. and Aebi,M. (2004) Roles of N-linked glycans in the endoplasmic reticulum. *Annu. Rev. Biochem.*, **73**, 1019–1049.
Huang,L.S. and Sternberg,P.W. (1995) Genetic dissection of developmental pathways. *Methods Cell Biol.*, **48**, 97–122.
Hughes,T.R. (2005) Universal epistasis analysis. *Nat. Genet.*, **37**, 457–457.
Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
Jaimovich,A. and Friedman,N. (2011) From large-scale assays to mechanistic insights: computational analysis of interactions. *Curr. Opin. Biotechnol.*, **22**, 87–93.
Jonikas,M.C. *et al.* (2009) Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science*, **323**, 1693–1697.
Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36 (Suppl. 1)**, D480–D484.
Kim,W. *et al.* (2005) Yos9p detects and targets misfolded glycoproteins for ER-associated degradation. *Mol. Cell*, **19**, 753–764.
Koren,Y. *et al.* (2009) Matrix factorization techniques for recommender systems. *Computer*, **42**, 30–37.
Lee,D.D. and Seung,H.S. (2000) Algorithms for non-negative matrix factorization. In: *Advances in Neural Information Processing Systems*. MIT Press, Denver, Colorado, pp. 556–562.
Mani,R. *et al.* (2008) Defining genetic interaction. *Proc. Natl Acad. Sci. USA*, **105**, 3461–3466.
Metzstein,M.M. *et al.* (1998) Genetics of programmed cell death in *C. elegans*: past, present and future. *Trends Genet.*, **14**, 410–416.
Min,J.H. and Lee,Y.C. (2005) Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Syst. Appl.*, **28**, 603–614.
Mohammadi,S. *et al.* (2012) Role of synthetic genetic interactions in understanding functional interactions among pathways. *Pac. Symp. Biocomput.*, **17**, 43–54.
Nakatsukasa,K. and Brodsky,J.L. (2008) The recognition and retrotranslocation of misfolded proteins from the endoplasmic reticulum. *Traffic*, **9**, 861–870.
Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
Phenix,H. *et al.* (2011) Quantitative epistasis analysis and pathway inference from genetic interaction data. *PLoS Comput. Biol.*, **7**, e1002048.
Phenix,H. *et al.* (2013) Identifiability and inference of pathway motifs by epistasis analysis. *Chaos*, **23**, 025103.
Qi,Y. *et al.* (2008) Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res.*, **18**, 1991–2004.
Romero,P.A. *et al.* (1997) The yeast CWH41 gene encodes glucosidase I. *Glycobiology*, **7**, 997–1004.
Roth,F.P. *et al.* (2009) Q&A: epistasis. *J. Biol.*, **8**, 35.
Schuldiner,M. *et al.* (2008) The GET complex mediates insertion of tail-anchored proteins into the ER membrane. *Cell*, **134**, 634–645.
St Onge,R.P. *et al.* (2007) Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat. Genet.*, **39**, 199–206.
Stefanovic,S. and Hegde,R.S. (2007) Identification of a targeting factor for post-translational membrane protein insertion into the ER. *Cell*, **128**, 1147–1159.
Surma,M.A. *et al.* (2013) A lipid E-MAP identifies Ubx2 as a critical regulator of lipid saturation and lipid bilayer stress. *Mol. Cell*, **51**, 519–530.
Tong,A.H. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
Van Driessche,N. *et al.* (2005) Epistasis analysis with global transcriptional phenotypes. *Nature Genetics*, **37**, 471–477.
Williams,C.K. and Rasmussen,C.E. (1996) Gaussian processes for regression. In: *Advances in Neural Information Processing Systems*. MIT Press, Denver, Colorado, pp. 514–520.
Zhang,X.D. and Zhao,X.M. (2013) Computational approaches for identifying signaling pathways from molecular interaction networks. *Curr. Bioinform.*, **8**, 56–62.
Zupan,B. *et al.* (2003) GenePath: a system for automated construction of genetic networks from mutant data. *Bioinformatics*, **19**, 383–389.