Imputation of Quantitative Genetic Interactions in Epistatic MAPs by Interaction Propagation Matrix Completion

Marinka Žitnik[1] and Blaž Zupan[1,2]

[1] Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25,
1000, Slovenia
[2] Department of Molecular and Human Genetics, Baylor College of Medicine,
Houston, TX-77030, USA
{marinka.zitnik,blaz.zupan}@fri.uni-lj.si

Abstract. A popular large-scale gene interaction discovery platform is the Epistatic Miniarray Profile (E-MAP). E-MAPs benefit from quantitative output, which makes it possible to detect subtle interactions. However, due to the limits of biotechnology, E-MAP studies fail to measure genetic interactions for up to 40% of gene pairs in an assay. Missing measurements can be recovered by computational techniques for data imputation, thus completing the interaction profiles and enabling downstream analysis algorithms that could otherwise be sensitive to largely incomplete data sets. We introduce a new interaction data imputation method called interaction propagation matrix completion (IP-MC). The core part of IP-MC is a low-rank (latent) probabilistic matrix completion approach that considers additional knowledge presented through a gene network. IP-MC assumes that interactions are transitive, such that latent gene interaction profiles depend on the profiles of their direct neighbors in a given gene network. As the IP-MC inference algorithm progresses, the latent interaction profiles propagate through the branches of the network. In a study with three different E-MAP data assays and the considered protein-protein interaction and Gene Ontology similarity networks, IP-MC significantly surpassed existing alternative techniques. Inclusion of information from gene networks also allows IP-MC to predict interactions for genes that were not included in original E-MAP assays, a task that could not be considered by current imputation approaches.

Keywords: genetic interaction, missing value imputation, epistatic miniarray profile, matrix completion, interaction propagation.

1 Introduction

The epistatic miniarray profile (E-MAP) technology [1–4] is based on a synthetic genetic array (SGA) approach [5,6] and generates quantitative measures of both positive and negative genetic interactions (GIs) between gene pairs. E-MAP was developed to study the phenomenon of epistasis, wherein the presence of

R. Sharan (Ed.): RECOMB 2014, LNBI 8394, pp. 448–462, 2014.

one mutation modulates the effect of another mutation. The power of epistasis analysis is greatly enhanced by quantitative GI scores [2]. E-MAP has provided high-throughput measurements of hundreds of thousands of GIs in yeast [1, 4, 7] and has been shown to significantly improve gene function prediction [7]. However, E-MAP data suffer from a large number of missing values, which can be as high as ∼40% for a given assay (see also Table 1). Missing values correspond to pairs of genes for which the strength of the interaction could not be measured during the experimental procedure or that were subsequently removed due to low reliability. A high proportion of missing values can adversely affect analysis algorithms or even prevent their use. For instance, missing data might introduce instability in clustering results [8] or bias the inference of prediction models [9]. Accurate imputation of quantitative GIs is therefore an appealing option to improve downstream data analysis and correspondence between genetic and functional similarity [7, 10–13]. Imputed quantitative GIs can be a powerful source for understanding both the functions of individual genes and the relationships between pathways in the cell.

The missing value problem in E-MAPs resembles that from gene expression data, where imputation has been well studied [9, 14, 15]. The objective in both tasks is to estimate the values of missing entries given the incomplete data matrix. Both types of data may exhibit a correlation between gene or mutant profiles, which is indicative of co-regulation in the case of gene expression data and pathway membership in the case of E-MAP data [16]. E-MAP data sets are therefore often analyzed with tools originally developed for gene expression data analysis [17]. However, there are important differences between E-MAP and gene expression data that limit the direct application of gene expression imputation techniques to E-MAPs [16]. E-MAP data are pairwise, symmetric and have substantially different dimensionality than gene expression data sets. They contain considerably more missing values than gene expression data sets (the latter have up to a 5% missing data rate, see [9, 18]). These differences, coupled with the biological significance of E-MAP studies, have spurred the development of specialized computational techniques for recovering missing data in E-MAP-like data sets [16].

In this paper, we propose IP-MC ("interaction propagation matrix completion"), a *hybrid* and *knowledge assisted* method for imputing missing values in E-MAP-like data sets. IP-MC builds upon two concepts, matrix completion and propagation of interaction. Matrix completion uses information on global correlation between entries in the E-MAP score matrix. The interaction propagation serves to exploit the local similarity of genes in a gene network. The use of background knowledge in the form of gene networks gives IP-MC the potential to improve imputation accuracy beyond purely data-driven approaches. This could be especially important for data sets with a small number of genes and a high missing data rate, such as E-MAPs. In the following, we derive a mathematical formulation of the proposed approach and, in a comparative study that includes several state-of-the-art imputation techniques, demonstrate its accuracy across several E-MAP data sets.

2 Related Work

Imputation algorithms for gene expression data sets are reviewed in Liew *et al.* (2011) [9], who categorized them into four classes based on how they utilize or combine local and global information from within the data (*local, global* and *hybrid* algorithms) and their use of domain knowledge during imputation (*knowledge-assisted* algorithms). Local methods, such as k-nearest neighbors (KNNimpute) [14], local least squares (LLS) [19] and adaptive least squares (LSimpute) [18], rely on the local similarity of genes to recover the missing values. Global methods are based on matrix decompositions, such as the singular value decomposition (e.g. SVDimpute [14]), the singular value thresholding algorithm for matrix completion (SVT) [20] and Bayesian principal component analysis (BPCA) [21]. A hybrid imputation approach for gene expression data by Jörnsten *et al.* (2005) [22] estimates missing values by combining estimates from three local and two global imputation methods.

Only a handful of missing data imputation algorithms directly address E-MAP-like data sets. Ulitsky *et al.* (2009) [23] experimented with a variety of genomic features, such as the existence of physical interaction or co-expression between gene pairs, that were used as input to a classification algorithm. The IP-MC differs from this approach as it directly uses the matrix of measured GI scores and does not require data-specific feature engineering. Ryan *et al.* (2010, 2011) [16, 24] considered four general strategies for imputing missing values – three local methods and one global method – and adapted these strategies to address E-MAPs. They modified unweighted and weighted k-nearest neighbors imputation methods (uKNN and wNN, respectively). They also adapted LLS and BPCA algorithms to handle symmetric data. We refer the reader to Ryan *et al.* (2010) [16] for details on the algorithm modifications. We compare their imputation approaches with the IP-MC (see Sec. 5). Pan *et al.* (2011) [25] proposed an ensemble approach to combine the outputs of two global and four local imputation methods based on diversity of estimates of individual algorithms. In this paper we focus on the development of a single algorithm that if necessary could be used in an ensemble, and therefore compare it only with ensemble-free algorithms.

Another avenue of research focuses on predicting qualitative, i.e. binary, instead of quantitative interactions. Qualitative predictions estimate the presence or absence of certain types of interaction rather than their strength [26–29]. A major distinction between these techniques and the method proposed in the paper is that we aim at accurate imputation of quantitative genetic interactions using the scale of GI scores. Individual GI may by itself already provide valuable biological insight, as each interaction provides evidence for a functional relationship between a gene pair. Prediction of synthetic sick and lethal interaction types in *S. cerevisiae* was pioneered by Wong *et al.* (2004) [26], who applied probabilistic decision trees to diverse genomic data. Wong *et al.* [26] introduced *2-hop features* for capturing the relationship between a gene pair and a third gene. They showed that, for example, if protein g_1 physically interacts with protein g_2, and gene g_3 is synthetic lethal with the encoding gene of g_2, then this increases the likelihood of a synthetic lethal interaction between the encoding gene of g_1 and gene g_3. Two-hop features were shown to be crucial when predicting GIs [11, 23, 26] and are the rationale behind our concept of interaction propagation.

3 Methods

We first introduce a probabilistic model of matrix completion for missing value imputation in E-MAP-like data sets. The model predicts scores for missing interaction measurements by employing only the E-MAP score matrix. We then extend it with the notion of interaction propagation. The resulting method, IP-MC, is able to exploit the transitivity of interactions, that is, the relationship between a gene pair and a third gene (see Sec. 2). IP-MC predicts missing values from both E-MAP data and the associated gene network that encodes domain knowledge. Any type of knowledge that can be expressed in the form of a network can be passed to IP-MC. In this paper, we use the Gene Ontology [30] semantic similarity network and protein-protein interaction network.

3.1 Problem Definition and Preliminaries

In the E-MAP study we have a set of genes $(g_1, g_2, \ldots, g_n)$. The genetic interaction between a pair of genes is scored according to the fitness of the corresponding double mutant and reported through an S-score that reflects the magnitude and sign of the observed GI [2]. Scored GIs are reported in the form of a partially observed matrix $\mathbf{G} \in \mathbf{R}^{n \times n}$. In this matrix, $\mathbf{G}_{i,j}$ contains a GI measurement between g_i and g_j. Here, $\mathbf{G}$ is symmetric, $\mathbf{G}_{i,j} = \mathbf{G}_{j,i}$. Without loss of generality, we map GIs to the [0,1]-interval by normalizing $\mathbf{G}$ (step 1 in Fig. 2). Following the imputation, we re-scale the completed (imputed) matrix $\widehat{\mathbf{G}}$ to the original scale of S-scores (step 5 in Fig. 2).

In a gene network every gene g_i has a set of N_{g_i} neighbors, and $\mathbf{P}_{i,j}$ denotes the value of influence that gene $g_j \in N_{g_i}$ has on g_i. These values are given in matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$. We normalize each row of $\mathbf{P}$ such that $\sum_{j=1}^{n} \mathbf{P}_{i,j} = 1$. A non-zero entry $\mathbf{P}_{i,j}$ denotes dependence of the g_i-th latent feature vector to the g_j-th latent feature vector. Using this idea, latent features of genes that are indirectly connected in the network become dependent after a certain number of algorithm steps, the number of steps being determined by the path distance between genes. Hence, information about gene latent representation propagates through the network.

The model inference task is defined as follows: given a pair of genes, g_i and g_j, for which $\mathbf{G}_{i,j}$ (and $\mathbf{G}_{j,i}$) is unknown, predict the quantitative GI between g_i and g_j using $\mathbf{G}$ and $\mathbf{P}$. We employ a probabilistic view of matrix completion to learn gene latent feature vectors. Let $\mathbf{F} \in \mathbb{R}^{k \times n}$ and $\mathbf{H} \in \mathbb{R}^{k \times n}$ be gene latent feature matrices with column vectors $\mathbf{F}_i$ and $\mathbf{H}_j$ representing k-dimensional gene-specific latent feature vectors of g_i and g_j, respectively. The goal is to learn these latent feature matrices and utilize them for missing value imputation in E-MAP-like data sets.

3.2 Matrix Completion Model

We start our derivation by formulating basic matrix completion approach for recovering missing values in $\mathbf{G}$ without considering the additional gene network. Throughout the paper, this approach is denoted by MC. In order to learn low-dimensional gene latent feature matrices $\mathbf{F}$ and $\mathbf{H}$, we factorize observed values in $\mathbf{G}$. The conditional probability of observed GIs is defined as:

$$p(\mathbf{G}|\mathbf{F},\mathbf{H},\sigma_{\mathbf{G}}^2) = \prod_{i=1}^{n}\prod_{j=1}^{n} \mathcal{N}(\mathbf{G}_{i,j}|g(\mathbf{F}_i^T\mathbf{H}_j),\sigma_{\mathbf{G}}^2)^{I_{i,j}^{\mathbf{G}}}, \tag{1}$$

where $\mathcal{N}(x|\mu,\sigma^2)$ is a normal distribution with mean μ and variance σ^2 and $I_{i,j}^{\mathbf{G}}$ is an indicator function that is equal to 1 if a GI score between g_i and g_j is available and is 0 otherwise. Notice that Eq. (1) deals only with observed entries in matrix $\mathbf{G}$. Thus, predictions are not biased by setting missing entries in $\mathbf{G}$ to some fixed value, which is otherwise common in matrix factorization algorithms. The function g is a logistic function, $g(x) = 1/(1+e^{-0.5x})$, which bounds the range of $g(\mathbf{F}_i^T\mathbf{H}_j)$ within interval $(0,1)$. We assume a zero-mean Gaussian prior for gene latent feature vectors in $\mathbf{F}$ as $p(\mathbf{F}|\sigma_{\mathbf{F}}^2) = \prod_{i=1}^{n}\mathcal{N}(\mathbf{F}_i|0,\sigma_{\mathbf{F}}^2\mathbf{I})$ and similarly, the prior probability distribution for $\mathbf{H}$ is given by $p(\mathbf{H}|\sigma_{\mathbf{H}}^2) = \prod_{i=1}^{n}\mathcal{N}(\mathbf{H}_i|0,\sigma_{\mathbf{H}}^2\mathbf{I})$.

Through Bayesian inference we obtain the following equation for the log-posterior probability of latent feature matrices $\mathbf{F}$ and $\mathbf{H}$ given the interaction measurements in $\mathbf{G}$:

$$\begin{aligned}\ln p(\mathbf{F},\mathbf{H}|\mathbf{G},\sigma_{\mathbf{G}}^2,\sigma_{\mathbf{F}}^2,\sigma_{\mathbf{H}}^2) = &-\frac{1}{2\sigma_{\mathbf{G}}^2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{i,j}^{\mathbf{G}}(\mathbf{G}_{i,j} - g(\mathbf{F}_i^T\mathbf{H}_j))^2 - \frac{1}{2\sigma_{\mathbf{F}}^2}\sum_{i=1}^{n}\mathbf{F}_i^T\mathbf{F}_i \\ &- \frac{1}{2\sigma_{\mathbf{H}}^2}\sum_{j=1}^{n}\mathbf{H}_j^T\mathbf{H}_j - \frac{1}{2}(\sum_{i=1}^{n}\sum_{j=1}^{n} I_{i,j}^{\mathbf{G}})\ln\sigma_{\mathbf{G}}^2 - \frac{1}{2}nk(\ln\sigma_{\mathbf{F}}^2 + \ln\sigma_{\mathbf{H}}^2) + \mathcal{C}. \end{aligned} \tag{2}$$

We select the factorized model by finding the maximum *a posteriori* (MAP) estimate. This is equivalent to solving a minimization problem with the objective:

$$\mathcal{L}(\mathbf{G},\mathbf{F},\mathbf{H}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{i,j}^{\mathbf{G}}(\mathbf{G}_{i,j} - g(\mathbf{F}_i^T\mathbf{H}_j))^2 + \frac{\lambda_{\mathbf{F}}}{2}\sum_{i=1}^{N}\mathbf{F}_i^T\mathbf{F}_i + \frac{\lambda_{\mathbf{H}}}{2}\sum_{j=1}^{N}\mathbf{H}_j^T\mathbf{H}_j, \tag{3}$$

where $\lambda_{\mathbf{F}} = \sigma_{\mathbf{G}}^2/\sigma_{\mathbf{F}}^2$ and $\lambda_{\mathbf{H}} = \sigma_{\mathbf{G}}^2/\sigma_{\mathbf{H}}^2$. Interactions in $\mathbf{G}$ are normalized before numerical optimization such that they are between 0 and 1 because their estimates $g(\mathbf{F}^T\mathbf{H})$ are also bounded. We keep the observation noise variance $\sigma_{\mathbf{G}}^2$ and prior variances $\sigma_{\mathbf{F}}^2$ and $\sigma_{\mathbf{H}}^2$ fixed and use a gradient descent algorithm to find the local minimum of $\mathcal{L}(\mathbf{G},\mathbf{F},\mathbf{H})$ to infer gene latent feature matrices.

3.3 Interaction Propagation Matrix Completion Model

Interaction propagation matrix completion (IP-MC) extends the basic matrix completion model MC by borrowing latent feature information from neighboring genes in the network $\mathbf{P}$. A graphical example of IP-MC is shown in Fig. 1.

The biological motivation for the propagation of interactions stems from the transitive relationship between a gene pair and a third gene (see Sec. 2) and indicates that the behavior of a gene is affected by its direct and indirect neighbors in the underlying gene network $\mathbf{P}$. In other words, the latent feature vector of gene g, $\mathbf{F}_g$, is in each iteration dependent on the latent feature vectors of its direct neighbors $h \in N_g$ in $\mathbf{P}$. The influence is formulated as $\widehat{\mathbf{F}}_g = \sum_{h \in N_g} \mathbf{P}_{g,h}\mathbf{F}_h$, where $\widehat{\mathbf{F}}_g$ is the estimated latent feature vector of g given feature vectors of its direct neighbors. Thus, the latent feature vectors in $\mathbf{F}$ of genes that are indirectly connected in network $\mathbf{P}$ are dependent and thus, information about their latent representation propagates as the algorithm progresses according to the connectivity of network $\mathbf{P}$.

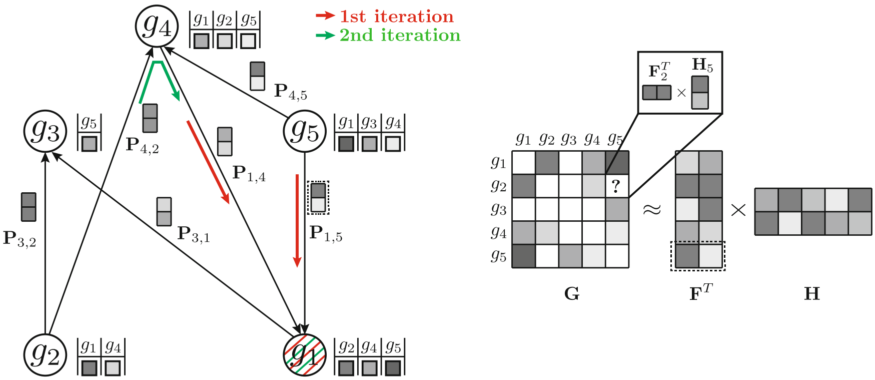


Fig. 1. An example application of the interaction propagation matrix completion algorithm (IP-MC). A hypothetical E-MAP data set with five genes ($g_1, \dots, g_5$) is given. Their measured GI profiles are listed next to corresponding nodes in gene network $\mathbf{P}$ (left) and are shown in the sparse and symmetric matrix $\mathbf{G}$ (right). Different shades of grey quantify interaction strength, while white matrix entries in $\mathbf{G}$ denote missing values. Matrices $\mathbf{F}$ and $\mathbf{H}$ are gene latent feature matrices. Gene latent feature vector $\mathbf{F}_{g_i}$ depends in each iteration of IP-MC on the latent feature vectors of g_i's direct neighbors in $\mathbf{P}$. For instance, the latent vector of gene g_1 in $\mathbf{F}$ depends in the first iteration of the IP-MC update (in red) only on its direct neighbors, the latent vectors of g_4 and g_5 ($\mathbf{F}_{g_4}$ and $\mathbf{F}_{g_5}$ are shown on input edges of g_1), whose level of influence is determined by $\mathbf{P}_{1,4}$ and $\mathbf{P}_{1,5}$, respectively. In the second iteration, the update of $\mathbf{F}_{g_1}$ (in green) also depends indirectly on the latent vector of g_2, $\mathbf{F}_{g_2}$. Thus, the influence of gene latent feature vectors propagates in $\mathbf{P}$. Gene latent feature matrix $\mathbf{H}$ is not influenced by the gene neighborhood in $\mathbf{P}$.

Notice that considering gene network $\mathbf{P}$ does not change the conditional probability of observed measurements (Eq. (1)). It only affects gene latent feature vectors in $\mathbf{F}$. We describe them with two factors: the zero-mean Gaussian prior to avoid overfitting and the conditional distribution of gene latent feature vectors given the latent feature vectors of their direct neighbors:

$$p(\mathbf{F}|\mathbf{P}, \sigma_{\mathbf{F}}^2, \sigma_{\mathbf{P}}^2) \propto \prod_{i=1}^{n} \mathcal{N}(\mathbf{F}_i|0, \sigma_{\mathbf{F}}^2\mathbf{I}) \times \prod_{i=1}^{n} \mathcal{N}(\mathbf{F}_i| \sum_{j \in N_i} \mathbf{P}_{i,j}\mathbf{F}_j, \sigma_{\mathbf{P}}^2\mathbf{I}). \tag{4}$$

Notice that such formulation of gene latent matrix $\mathbf{F}$ keeps gene feature vectors $\mathbf{F}_i$ both small and close to the latent feature vectors of their direct neighbors. Much like the previous section, we get the following equation through Bayesian inference for the posterior probability of gene latent feature matrices $\mathbf{F}$ and $\mathbf{H}$ given observed GI scores $\mathbf{G}$ and gene network $\mathbf{P}$:

$$\begin{aligned} p(\mathbf{F}, \mathbf{H}|\mathbf{G}, \mathbf{P}, \sigma_{\mathbf{G}}^2, \sigma_{\mathbf{P}}^2, \sigma_{\mathbf{F}}^2, \sigma_{\mathbf{H}}^2) \propto & \prod_{i=1}^{n}\prod_{j=1}^{n} \mathcal{N}(\mathbf{G}_{i,j}|g(\mathbf{F}_i^T\mathbf{H}_j), \sigma_{\mathbf{G}}^2)^{I_{i,j}^{\mathbf{G}}} \\ & \times \prod_{i=1}^{n} \mathcal{N}(\mathbf{F}_i| \sum_{j \in N_i} \mathbf{P}_{i,j}\mathbf{F}_j, \sigma_P^2\mathbf{I}) \times \prod_{i=1}^{n} \mathcal{N}(\mathbf{F}_i|0, \sigma_{\mathbf{F}}^2\mathbf{I}) \times \prod_{j=1}^{n} \mathcal{N}(\mathbf{H}_j|0, \sigma_{\mathbf{H}}^2\mathbf{I}). \end{aligned} \tag{5}$$

We then compute the log-posterior probability to obtain an equation similar to Eq. (2) but with an additional term due to the interaction propagation concept. To maximize conditional posterior probability over gene latent features, we fix the prior and observation noise variance and employ gradient descent on $\mathbf{F}$ and $\mathbf{H}$. In particular, we minimize the objective function similar to Eq. (3) that has an additional term to account for the conditional probability of gene latent features given their neighborhoods in gene network $\mathbf{P}$. The complete algorithm of IP-MC is presented in Fig. 2. In each iteration, gene latent feature matrices $\mathbf{F}$ and $\mathbf{H}$ are updated based on the latent feature vectors from the previous iteration and network neighborhood in $\mathbf{P}$. Successive updates of $\mathbf{F}_i$ and $\mathbf{H}_j$ converge to a maximum *a posteriori* (MAP) estimate of the posterior probability in Eq. (5).

4 Experimental Setup

In the experiments we consider an existing incomplete E-MAP matrix and artificially introduce an additional 1% of missing values for a set of arbitrarily selected gene pairs [16, 25]. These gene pairs and their data constitute a test set on which we evaluate the performance of imputation algorithms. Because of E-MAP symmetry, for a given test gene pair and its corresponding entry $\mathbf{G}_{i,j}$, we also hide the value of $\mathbf{G}_{j,i}$. We repeat this process 30 times and report on the averaged imputation performance.

Notice that the standard performance evaluation procedure of missing value imputation methods for gene expression data is not directly applicable to E-MAPs for the several reasons discussed in [16]. This approach constructs a complete gene expression data matrix by removing genes with missing data and then artificially introduces missing values for evaluation. In gene expression data, a substantially lower fraction of data is missing than in E-MAPs (Table 1)

```
Input: Sparse matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ containing S-scores of measured E-MAP interactions,
gene network $\mathbf{P} \in \mathbb{R}^{n \times n}$ and parameters $\lambda_{\mathbf{F}} = \lambda_{\mathbf{H}}$, $\lambda_{\mathbf{P}}$, rank $k$ and learning rate $\alpha$.
Output: Completed E-MAP matrix $\widehat{\mathbf{G}}$.

1. Normalize $\tilde{\mathbf{G}} = (\mathbf{G} - \min_{i,j} \mathbf{G}_{i,j}) / \max_{i,j} \mathbf{G}_{i,j}$.
2. Normalize each row of $\mathbf{P}$ such that $\sum_{j=1}^{n} \mathbf{P}_{i,j} = 1$.
3. Sample $\mathbf{F} \sim \mathcal{U}[0,1]^{k \times n}$ and $\mathbf{H} \sim \mathcal{U}[0,1]^{k \times n}$.
4. Repeat until convergence:
   • For $i, j \in 1, 2, \ldots, n$ :
         $\frac{\partial \mathcal{L}}{\partial \mathbf{F}_i} = \sum_{j=1}^{n} I_{i,j}^{\tilde{\mathbf{G}}} \mathbf{H}_j g'(\mathbf{F}_i^T \mathbf{H}_j)(g(\mathbf{F}_i^T \mathbf{H}_j) - \tilde{\mathbf{G}}_{i,j}) + \lambda_{\mathbf{F}} \mathbf{F}_i +$
             $+\lambda_{\mathbf{P}}(\mathbf{F}_i - \sum_{j \in N_i} \mathbf{P}_{i,j} \mathbf{F}_j) - \lambda_{\mathbf{P}} \sum_{\{j | i \in N_j\}} \mathbf{P}_{j,i}(\mathbf{F}_j - \sum_{l \in N_j} \mathbf{P}_{j,l} \mathbf{F}_l),$
         $\frac{\partial \mathcal{L}}{\partial \mathbf{H}_j} = \sum_{i=1}^{n} I_{i,j}^{\tilde{\mathbf{G}}} \mathbf{F}_i g'(\mathbf{F}_i^T \mathbf{H}_j)(g(\mathbf{F}_i^T \mathbf{H}_j) - \tilde{\mathbf{G}}_{i,j}) + \lambda_{\mathbf{H}} \mathbf{H}_j.$
   • Set $\mathbf{F}_i \leftarrow \mathbf{F}_i - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{F}_i}$ for $i = 1, 2, \ldots, n$.
   • Set $\mathbf{H}_j \leftarrow \mathbf{H}_j - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{H}_j}$ for $j = 1, 2, \ldots, n$.
5. Compute $\widehat{\mathbf{G}} = g(\mathbf{F}^T \mathbf{H}) \cdot \max_{i,j} \mathbf{G}_{i,j} + \min_{i,j} \mathbf{G}_{i,j}$. Impute missing entry $(i, j)$ as
   $(\widehat{\mathbf{G}}_{i,j} + \widehat{\mathbf{G}}_{j,i})/2$.
```

Fig. 2. Interaction propagation matrix completion (IP-MC) algorithm. We observed that parameter values $\lambda_{\mathbf{H}} = \lambda_{\mathbf{F}} = 0.01$ and $\alpha = 0.1$ gave accurate results across a number of different data sets. Parameter $\lambda_{\mathbf{P}}$, which controls the influence of gene network $\mathbf{P}$ on gene latent feature vectors $\mathbf{F}_i$, depended on data set complexity [15]. In data sets with higher complexity, we used a larger $\lambda_{\mathbf{P}}$ ($\lambda_{\mathbf{P}} = 1$).

and removing a small number of genes and experimental conditions does not significantly reduce the size of the data set.

In our experiments we select the number of latent dimensions k and regularization parameters $\lambda_{\mathbf{F}}$ and $\lambda_{\mathbf{P}}$ of IP-MC with the following procedure: For each data set and before the performance evaluation, we leave out 1% of randomly selected known values and attempt to impute them with varying values of parameters in a grid search fashion. Parameter values that result in the best estimation of the left-out values are then used in all experiments involving the data set. Notice that the left-out values are determined before the performance evaluation and are therefore not included in the test data set.

We consider two measures of imputation accuracy. These are the Pearson correlation (CC) between the imputed and the true values, and the normalized root mean squared error (NRMSE) [21] given as NRMSE $= \sqrt{\mathbb{E}((\widehat{\mathbf{y}} - \mathbf{y})^2)/\mathrm{Var}(\mathbf{y})}$, where $\mathbf{y}$ and $\widehat{\mathbf{y}}$ denote vectors of true and imputed values, respectively. More accurate imputations give a higher correlation score and a lower NRMSE.

To test if the differences in performance between imputation methods are significant, we use the Wilcoxon signed-rank test, a non-parametric equivalent of a paired t-test. Its advantage is that it does not require a normal distribution or homogeneity of variance, but it has less statistical power, so there is the risk that some differences are not recognized as significant.

5 Results and Discussion

We considered three E-MAP data sets and compared IP-MC to five state-of-the-art methods for imputing missing values in E-MAP-like data sets [16]. We set the parameters of these methods to values as proposed in [16] (wNN, LLS, BPCA) or optimized the parameter selection through a grid search (SVT, MC, IP-MC). The evaluated data sets are from the budding yeast *S. cerevisiae*; they differ in their size, the subset of genes that are studied and the proportion of missing values (Table 1). We used GI S-scores reported in original publications:

- Chromosome Biology [7]: This is the largest of the E-MAPs, encompassing interactions between 743 genes involved in various aspects of chromosome biology, such as chromatid segregation, DNA replication and transcriptional regulation.
- RNA processing (RNA) [4]: It focuses on the relationships between and within RNA processing pathways involving 552 mutations, 166 of which are hypomorphic alleles of essential genes.
- The Early Secretory Pathway (ESP) [1]: It generates genetic interaction maps of genes acting in the yeast early secretory pathway to identify pathway organization and components of physical complexes.

Table 1. Overview of the E-MAPs considered

Data set	Genes	Missing Interactions	Measured Interactions
Chromosome Biology [7]	743	34.0%	187,000
Early Secretory Pathway [1]	424	7.5%	83,000
RNA [4]	552	29.6%	107,000

IP-MC considered two different data sources for gene network **P**. The first network was constructed from Gene Ontology [30] (GO) annotation data as a weighted network of genes included in the E-MAP study in which edge weights corresponded to the number of shared GO terms between connected genes, excluding annotations inferred from GI studies (i.e. those with the IGI evidence code). The second network represented physical interaction data from BioGRID 3.2 [31]. The physical interaction network was a binary network in which two genes were connected if their gene products physically interact. Both networks were normalized as described in Sec. 3.1. Depending on a network, we denote their corresponding IP-MC models by IP-MC-GO and IP-MC-PPI, respectively.

5.1 Imputation Performance

Table 2 shows the CC and NRMSE scores of imputation algorithms along with the baseline method of filling-in with zeros. IP-MC-PPI and IP-MC-GO demonstrated the best accuracy on all considered data sets. We compared their scores with the performance of the second-best method (i.e. LLS on Chromosome Biology data set, SVT on ESP data set and MC on RNA data set) and found that improvements were significant in all data sets.

We did not observe any apparent connection between the proportion of missing values in a data set and the performance of any of the imputation methods. The performance was better on smaller ESP and RNA data sets, although differences were small and further investigation appears to be worthwhile.

The baseline method of filling-in with zeros had the worst performance for all data sets. While this approach seems naïve, it is justified by the expectation that most genes do not interact. We observed that BPCA failed to match the performance of weighted neighbor-based and local least squares methods, wNN and LLS, respectively, despite BPCA being an improvement of the KNN algorithm. Both local imputation methods, wNN and LLS, demonstrated good performance across all three data sets. The good performance of neighbor-based methods on larger data sets could be explained by a larger number of neighbors to choose from when imputing missing values, which resulted in more reliable missing value estimates.

Global methods, BPCA, SVT and MC, performed well on the ESP data set but poorly on the much larger Chromosome Biology data set. These methods assume the existence of a global covariance structure among all genes in the E-MAP score matrix. When this assumption is not appropriate, i.e. when the genes exhibit dominant local similarity structures, their imputation becomes less accurate. Notice that the comparable performance of SVT and MC across data sets was expected. Both methods solve related optimization problems and operate under the assumption that the underlying matrix of E-MAP scores is low-rank.

The superior performance of IP-MC models over other imputation methods can be explained by their ability to include circumstantial evidence. As a hybrid imputation approach, IP-MC can benefit from both global information present in E-MAP data and local similarity structure between genes. One could vary the level of influence of global and local imputation aspects on the inferred IP-MC model through the $\lambda_{\mathbf{P}}$ parameter, where a higher value of $\lambda_{\mathbf{P}}$ indicates more emphasis on locality. In this way, our approach can adequately address the data of varying underlying complexity [15], where the complexity denotes the difficulty with which the data can be mapped to a lower dimensional subspace. Brock *et al.* (2008) [15] devised an entropy-based imputation algorithm selection scheme based on their observation that global imputation methods performed better on gene expression data with lower complexity and that local methods performed better on data with higher complexity. Thus, their selection scheme could be adapted to work with E-MAP-like data sets and be used to set $\lambda_{\mathbf{P}}$ in an informed way, which is left for our future work. In additional experiments (results not shown), we found that the performance of IP-MC is robust for a broad range of $\lambda_{\mathbf{P}}$ parameter values.

Table 2. Accuracy as measured by the Pearson correlation coefficient (CC) and normalized root mean squared error (NRMSE) across three E-MAP data sets and eight imputation methods. MC denotes the matrix completion model (Sec. 3.2). The IP-MC-GO and IP-MC-PPI models are interaction propagation matrix completion models (Sec. 3.3) that utilize annotation and physical interaction data, respectively. For descriptions of other methods see Related Work. Highlighted results are significantly better than the best non-IP-MC method according to the Wilcoxon signed-rank test at 0.05 significance level.

Approach	Chromosome Biology		ESP		RNA	
	CC	NRMSE	CC	NRMSE	CC	NRMSE
Filling with zeros	0.000	1.021	0.000	1.011	0.000	1.000
BPCA ($k = 300$)	0.539	0.834	0.619	0.796	0.589	0.804
wNN ($k = 50$)	0.657	0.744	0.625	0.776	0.626	0.787
LLS ($k = 20$)	0.678	0.736	0.626	0.764	0.626	0.776
SVT ($k = 40$)	0.631	0.753	0.672	0.719	0.649	0.765
MC ($k = 40$)	0.641	0.742	0.653	0.722	0.651	0.760
IP-MC-GO ($k = 60$)	0.691	0.693	**0.732**	**0.648**	**0.727**	**0.641**
IP-MC-PPI ($k = 60$)	**0.722**	**0.668**	**0.742**	0.667	0.701	**0.652**

5.2 Missing Value Abundance and Distribution

Ulitsky *et al.* (2009) [23] described three different scenarios of missing values in E-MAP experiments (Fig. 3). The simplest and the most studied scenario is the *Random* model, for which we assume that missing measurements are generated independently and uniformly by some random process. The *Submatrix* model corresponds to the case when all interactions between a subset of genes (e.g. essential genes) are missing. The *Cross* model arises when all interactions between two disjoint subsets of genes are missing. This model concurs with the situation when two E-MAP data sets that share a subset of genes are combined into a single larger data set. We identify another missing value configuration, which we call the *Prediction* scenario (Fig. 4d). It occurs when GI profiles of a subset of genes are completely missing. Learning in such a setting is substantially harder as these genes do not have any associated measurements. In the previous section, we compared the imputation methods on the Random configuration, and study other configurations in this section. This time we were interested in the effect these configurations have on IP-MC, and we compared the algorithm to its variant MC that does not use additional knowledge (e.g. the gene network).

Fig. 4 reports the predictive performance of our matrix completion approach obtained by varying the fraction of missing values in the four missing data scenarios from Fig. 3. For $x = 5, 10, 20, \ldots, 90$ we hid $x\%$ of E-MAP measurements in ESP data and inferred prediction models. Our results are reasonably accurate (CC > 0.4) when up to 60% of the E-MAP values were hidden for the Random and Submatrix model. Notice that when we hid 60% of the ESP E-MAP measurements, the E-MAP scores were present in less than 40% of the matrix because the original ESP data set already had ~8% missing values (Table 1).

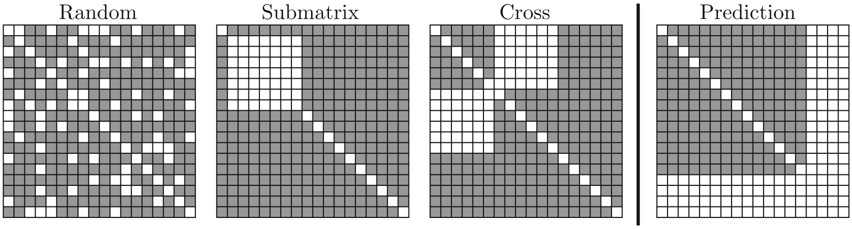
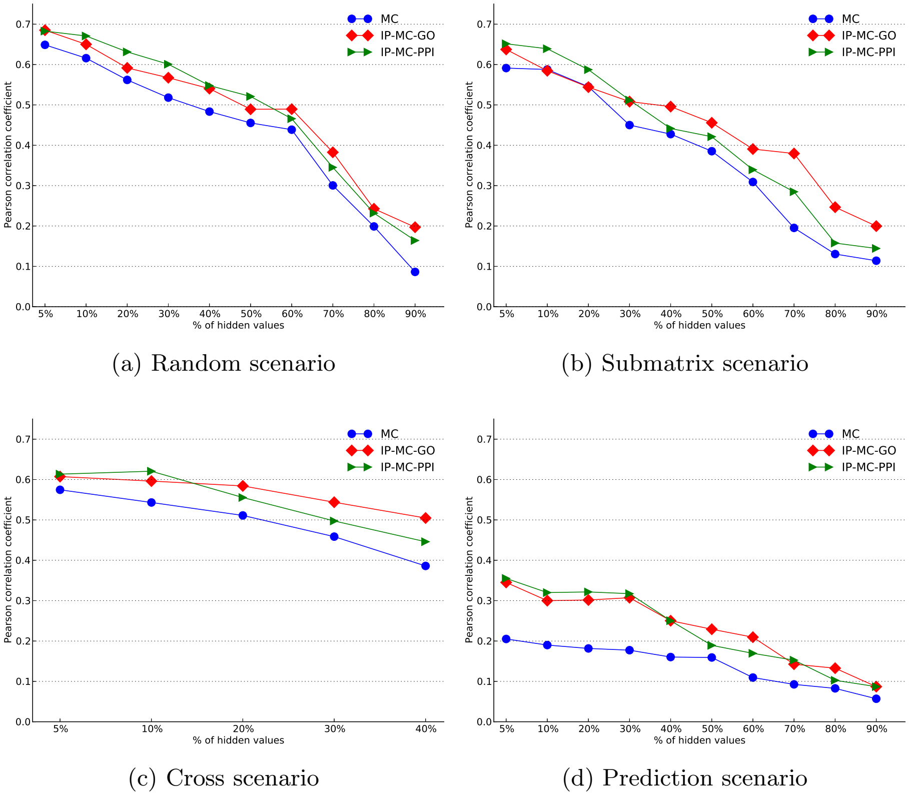


Fig. 3. The four configurations producing missing values in E-MAP data. In Random configuration, a random subset of GIs is hidden. In Submatrix or Cross configurations all interactions between a random subset of genes or two random disjoint subsets of genes, respectively, are hidden. In the Prediction scenario, complete profiles of GIs of a random subset of genes are removed.

When more than 80% of the data were removed, the three considered models still achieved higher accuracy (CC ≈ 0.2) than filling-in with zeros. As expected, predictions were more accurate for the Random model than for the Submatrix model for almost all fractions of hidden data (cf. Fig. 4a and Fig. 4b). However, the difference in performance between the Random and the Submatrix models tended to be small when less than 30% or more than 70% of the measurements were hidden. We observed that inclusion of additional genomic data is more useful in structured missing value scenarios, i.e. the Submatrix and Cross models (Fig. 4b–4c).

Imputation accuracy improved (Fig. 4) when E-MAP data were combined with gene annotation (IP-MC-GO) or protein-protein interaction (IP-MC-PPI) networks. These results are not surprising as several studies [6,7,32] showed that if two proteins act together to carry out a common function, deletions of the two encoding genes have similar GI profiles and that gene annotations from the GO and synthetic lethality are correlated, with 12 and 27% of genetic interactions having an identical or similar GO annotation, respectively [6]. Thus, our IP-MC-GO and IP-MC-PPI models could exploit the strong links between functionally similar genes, physically interacting proteins and GIs. The performance of our two integrated models indicates the importance of combining interaction and functional networks for predicting missing values in E-MAP data sets.

Imputation accuracy deteriorated when complete profiles of GIs were removed and IP-MC could only utilize circumstantial evidence (Fig. 4d). This suggests that measured gene pairs in the E-MAP are the best source of information for predicting missing pairs. However, as the percentage of missing GIs increases, the inclusion of other genomic data is more helpful. With the exception of the Prediction model, for which we observed the opposite behavior, the performance difference between MC and IP-MC was small ($\sim$10%) as long as $<$50% of the data were removed, but rose to above 20% when $\geq$60% of the data were removed (Fig. 4).



(a) Random scenario

(b) Submatrix scenario

(c) Cross scenario

(d) Prediction scenario

Fig. 4. Performance of imputation methods (Pearson correlation coefficient, CC) proposed in this paper for different fractions of missing values and scenarios of missing value distribution. Refer to the main text and Fig. 3 for descriptions of the missing value scenarios. MC denotes the matrix completion approach (Sec. 3.2). The integrated approaches are represented by IP-MC-GO and IP-MC-PPI (Sec. 3.3). Performance was assessed for the ESP E-MAP data set because it contains the least missing values. The 'Cross' configuration is not applicable when more than 50% of values are missing.

6 Conclusion

We have proposed a new missing value imputation method IP-MC that targets gene interaction data sets. The approach is unique in combining gene interaction and network data through inference of a single probabilistic model. Experiments on epistatic MAP interaction data sets show that the inclusion of additional knowledge is crucial and helps IP-MC to perform better than a number of state-of-the-art algorithms we have included in our study. The results are encouraging, have a potentially high practical value, and were intuitively expected. Gene interaction studies use double-mutant phenotypes to uncover functional dependencies, and additional knowledge that could provide any information on relations between genes should help. Driven by this intuition, the principal novelty of the paper is thus a new knowledge-based missing value imputation approach and the demonstration of its successful application on E-MAP data sets.

Acknowledgments. This work was supported by the Slovenian Research Agency (P2-0209, J2-5480), the National Institute of Health (P01-HD39691) and the European Commission (Health-F5-2010-242038).

References

1. Schuldiner, M., et al.: Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell 123(3), 507–519 (2005)
2. Collins, S.R., et al.: A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biology 7, R63 (2006)
3. Roguev, A., et al.: Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. Science 322(5900), 405–410 (2008)
4. Wilmes, G.M., et al.: A genetic interaction map of RNA-processing factors reveals links between Sem1/Dss1-containing complexes and mRNA export and splicing. Molecular Cell 32(5), 735–746 (2008)
5. Tong, A.H.Y., et al.: Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294(5550), 2364–2368 (2001)
6. Tong, A.H.Y., et al.: Global mapping of the yeast genetic interaction network. Science 303(5659), 808–813 (2004)
7. Collins, S.R., et al.: Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446(7137), 806–810 (2007)
8. de Brevern, A.G., et al.: Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. BMC Bioinformatics 5(1), 114 (2004)
9. Liew, A.W.C., et al.: Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics 12(5), 498–513 (2011)
10. Pu, S., et al.: Local coherence in genetic interaction patterns reveals prevalent functional versatility. Bioinformatics 24(20), 2376–2383 (2008)
11. Bandyopadhyay, S., et al.: Functional maps of protein complexes from quantitative genetic interaction data. PLoS Computational Biology 4(4), e1000065 (2008)
12. Ulitsky, I., et al.: From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. Molecular Systems Biology 4(1) (2008)
13. Järvinen, A.P., et al.: Predicting quantitative genetic interactions by means of sequential matrix approximation. PLoS One 3(9), e3284 (2008)
14. Troyanskaya, O., et al.: Missing value estimation methods for DNA microarrays. Bioinformatics 17(6), 520–525 (2001)
15. Brock, G.N., et al.: Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinformatics 9(1), 12 (2008)
16. Ryan, C., et al.: Missing value imputation for epistatic MAPs. BMC Bioinformatics 11(1), 197 (2010)

17. Zheng, J., et al.: Epistatic relationships reveal the functional organization of yeast transcription factors. Molecular Systems Biology 6(1) (2010)
18. Bø, T.H., et al.: LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Research 32(3), e34 (2004)
19. Kim, H., et al.: Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 21(2), 187–198 (2005)
20. Cai, J.F., et al.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization 20(4), 1956–1982 (2010)
21. Oba, S., et al.: A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19(16), 2088–2096 (2003)
22. Jörnsten, R., et al.: DNA microarray data imputation and significance analysis of differential expression. Bioinformatics 21(22), 4155–4161 (2005)
23. Ulitsky, I., et al.: Towards accurate imputation of quantitative genetic interactions. Genome Biology 10(12), R140 (2009)
24. Ryan, C., et al.: Imputing and predicting quantitative genetic interactions in epistatic MAPs. In: Network Biology, pp. 353–361 (2011)
25. Pan, X.Y., Tian, Y., Huang, Y., Shen, H.B.: Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach. Genomics 97(5), 257–264 (2011)
26. Wong, S.L., et al.: Combining biological networks to predict genetic interactions. PNAS 101(44), 15682–15687 (2004)
27. Kelley, R., Ideker, T.: Systematic interpretation of genetic interactions using protein networks. Nature Biotechnology 23(5), 561–566 (2005)
28. Qi, Y., et al.: 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 Research 18(12), 1991–2004 (2008)
29. Pandey, G., et al.: An integrative multi-network and multi-classifier approach to predict genetic interactions. PLoS Computational Biology 6(9), e1000928 (2010)
30. Ashburner, M., et al.: Gene Ontology: tool for the unification of biology. Nature Genetics 25(1), 25–29 (2000)
31. Stark, C., et al.: BioGRID: a general repository for interaction datasets. Nucleic Acids Research 34(suppl. 1), D535–D539 (2006)
32. Costanzo, M., et al.: The genetic landscape of a cell. Science 327(5964), 425–431 (2010)