Imputing Missing Values for Genetic Interaction Data

Yishu Wang, Lin Wang, Dejie Yang, Minghua Deng[1,1,1,1]

[a]Center for Quantitative Biology, Peking University,Beijing 100871
[b]Institute of Computing Technology, Chinese Academy of Science, Beijing100190
[c]School of Mathematical Sciences, Peking University, Beijing 100871
[d]Center for Statistical Sciences, Peking University, Beijing 100871

Abstract

Backgroud: Epistatic Miniarray Profiles (EMAP) enables the research of genetic interaction as an important method to construct large-scale genetic interaction network. However, high proportion of missing values frequently pose problems in the EMAP data analysis since they hinder useful information in the datasets. While there have been some imputation approaches available to EMAP data, we adopted an improved SVD modeling procedure to impute the missing values in EMAP data, which results in a higher accuracy rate comparing to existent methods.

Results: The improved SVD imputation method adopting an effective soft-threshold to SVD approach which has been showed to be the best model to impute the genetic interaction data comparing a number of advanced imputation methods. Imputation methods can also improve the result of clustering on EMAP dataset, where more meaningful modules, known pathways and protein complexes could be detected after applying our imputation method on EMAP dataset.

Conclusion: The results demonstrate that, while missing data are complicating unavoidably in EMAP data, we can complete the original dataset by the Soft-SVD approach to accurately recover the genetic interactions.

Keywords: Soft-SVD imputation EMAP

1. Introduction

Genetic interactions refer to the phenomenon whereby the mutation phenotype of two genes differs to the superimposition effect of two single mutations [1]. In budding yeast and fission yeast, genetic interactions can be acquired using high-throughput technologies Epistatic Miniarray Profile (EMAP) platform [2]. EMAP experiment construct double deletion strains systematically, by crossing query strains with a library of test strains. Then measure the colony size of the double mutant stains to get the S score which can indicate the genetic interaction: synthetic sick/lethal or alleviating [3]. In EMAP dataset, each gene in the query (or library) has its genetic interaction spectrum constructing by the genetic interaction S score with other genes in the library (or query). Researchers could exploit biology pathways and protein complexes by clustering the S score matrix.

However, one common characteristic of EMAP experiment is the significant high propotion of missing values - even up to 35% - that can reduce the effectiveness of the data analysis techniques such as cluster analysis, even prevent the use of some matrix factorization techniques such as SVD or PCA. One reason leading to the missing entries is that the genetic interaction strengths could not be measured by the high-throughput technologies, in addition some genetic interactions would be subsequently filtered due to unreliability.

The problem of missing values in genetic interaction datasets has been discussed before, but few technologies are used to impute quantitative epistasis values in EMAPs [4]. In previous papers, people improved some techniques used by imputing values in gene expression datasets and apply them in EMAP data. Four general strategies are considered in EMAP data - three local methods (nearest neighbor-based) and one global method. Previously, in order to improve the accuracy of missing value predictions, symmetric characters of datasets are involved into these original imputing techniques[4][5]. However, with the recent development of EMAP technology and the need of practical application, most of EMAP datasets are asymmetric[6][7][8], so symmetric characteristic could not be used in predicting missing entries on an increasing number of new EMAP experiment datasets. We extended original SVD method by giving it a soft threshold and some changes to optimization functions and restricted conditions [9]. This method which can be called Soft-SVD has been used in “Netflix” competition [9], image recovery [9] and eQTL [10], and has been demonstrated as a best efficient algorithm in these fields. Soft-SVD algorithm is not restricted by symmetry, so it can be used in widespread EMAP datasets. We introduced this methodology into imputing missing values in EMAP datasets, and we call it Soft-Impute in the following.

In this paper, we systematically stated how Soft-Impute method applied on EMAP datasets imputation, and then we conducted a detailed comparison of our method with other general imputing techniques, showing the marked improvement in estimation performance. Beyond the imputation accuracy, we evaluated the methods in terms of their ability to detect genetic interaction modules in which genes have similar interaction profiles and involved in the same physical complex or pathway, enriched in GO terms. We demonstrate that after imputing missing entries in EMAP score matrix, the downstream analysis: hierarchical clustering results are highly improved, and more significant genetic interaction modules can be exploited, which are enriched in the known discoveries.

The Soft-Impute methodology adopts the soft threshold to SVD algorithm, and proposes the nuclear norm to result in a convex optimization problem. It takes advantage of the relevance in dataset to impute missing entries. This algorithm is suitable to the dataset where there are modules in which entries are with high correlation. In this paper we constructed a synthetic dataset with low rank to test the effect of Soft-Impute and compare the imputation accuracy of different imputation methods. EMAP datasets store genetic interaction spectrums, where genes in the same protein complex or biology pathway tend to have the similar genetic interaction spectrums. So in EMAP data matrixes there are several modules with high correlation. As a whole, the matrix is low rank which satisfy the request of Soft-Impute algorithm for datasets. We apply the proposed method to three public EMAP datasets.

The missing value problem is not new in genetic interaction datasets. We introduced an appropriate methodology: Soft-Impute to solve this problem and to promote the efficiency of downstream data analysis at the same time.

2. Materials

2.1. Epistatic miniarray profile (EMAP)

The genetic interaction datasets here we used are from EMAP analysis. Three EMAP datasets are used in our analysis, including a dataset studying the early secretory pathway (ESP) [11], and a dataset studying the chromosome biology (CHR) [12] for budding yeast, as well as a genome-wide EMAP profile for fission yeast [13]. The first two datasets are symmetric matrix, which query genes are the same as library genes. There are 424 genes with about 80000 pairwise measurements in ESP dataset, containing about 7.5% missing entries, while there are 743 genes with about 200000 pairwise measurements in CHR dataset, containing about 34% missing entries. On the contrary, the genome-scale genetic interaction matrix of Fission yeast is not symmetric, containing 953 alleles of 876 genes against a mutant library of more than 2000 deletions, resulting in an EMAP profile with 1.6 million genetic interactions, with 16% missing entries.

2.2. Synthetic data

We created a synthetic dataset with low rank to realize Soft-Impute algorithm and compare it with other imputation methods. We assume there are k modules in synthetic dataset, in which entries in the same module are with higher relevance, on the contrary, entries not in the same module are with lower relevance. We constructed a dataset of 250 elements representing query genes in 500 dimension standing for library genes as following:

1. We first construct a vector of 500 elements randomly chosen from ESP dataset. Denoted by $\overrightarrow{A_1} = \{a_1, a_2, ..., a_{500}\}$

(a) Multiply by Gaussian noise to every element a_i of $\overrightarrow{A_1}$, which is randomly chosen from $N(1, |a_i|)$. Denoted by $\overrightarrow{A_2}$

(b) Repeat (a) for k times and result in k vectors $\{\overrightarrow{A}_1, ...\overrightarrow{A}_k\}$

2. Generate a matrix with rank k using above k vectors.

(a) Generate k random numbers $n_1, n_2, ..., n_k$, ranging from 3 to 30, such that their sum is 250.

(b) Generate a matrix with 250 vectors by repeating vector $\overrightarrow{A_1}$ for n_1 times, vector $\overrightarrow{A_2}$ for n_2 times,..., vector $\overrightarrow{A_k}$ for n_k times. Apparently, such a matrix is of rank k.

3. Add a Gaussian noise drawn from $N(0, 0.5)$ to each entry of the above matrix.

4. Now we construct a matrix of 250 vectors with 500 dimensions.

3. Model and Algorithm

3.1. Soft-Impute Model

EMAP data can be represented by a matrix $X_{m \times n}$, where m and n represent the number of query genes and library genes. As the existing of missing entries in EMAP dataset matrix, $\Omega \subset \{1, ..., m\} \times \{1, ..., n\}$ denotes the indices of observed entries. So $\mathbf{X}$ is the original data with observed entries denoted by Ω and missing values denoted by $\Omega^{\perp}$. To impute this matrix, we aim to find a complete matrix $\mathbf{Z}$, which is close to $\mathbf{X}$ on the observed entries Ω, and has a low rank. Here low rank assumption is based on the consideration that the genetic interaction profile for co-functional genes are shown to have high correlation relationships [11][12]. N. Srebro et al. have studied generalization error bounds for learning the low-rank matrices [14], It is also showed theoretically that the true underlying matrix can be recovered within very high accuracy under certain assumptions on the entries of the matrix, locations, and proportion of unobserved entries [15][16][17]. R. Mazumder et al. formulate the above problem as the following optimization problem [9]:

$$\begin{aligned} &\textit{minimize} \;\; \textit{rank}(\mathbf{Z}) \\ &\textit{subject to} \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2 \leqslant \delta \end{aligned} \tag{1}$$

Where $\delta \geq 0$ is a regularization parameter to control the error tolerance.

However in the optimization above, the rank constraint makes the problem combinatorially hard for general Ω [18]. One small modification to (1) is [9]:

$$\begin{aligned} &\textit{minimize} \;\; \| \mathbf{Z} \|_* \\ &\textit{subject to} \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2 \leqslant \delta \end{aligned} \tag{2}$$

Where $\| \mathbf{Z} \|_*$ is the nuclear norm of $\mathbf{Z}$ ($\|\mathbf{Z}\|_* = \sum_{i=1}^{r} \sigma_i$, where $\sigma_1, ..., \sigma_r$ are the singular values of $\mathbf{Z}$ and r is the rank of $\mathbf{Z}$). This modification makes problem convex [19]. Such a problem can be reformulated (2) to the Lagrange form [9]:

$$\underset{Z}{\textit{minimize}} \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2 + \lambda \|\mathbf{Z}\|_* \tag{3}$$

Here $\lambda \geq 0$ is a regularization parameter controlling the nuclear norm of estimated value $\widetilde{\mathbf{Z}}_\lambda$ of (3).

Suppose we only observed a subset of $\mathbf{X}$, indexed by Ω, and the missing entries are indexed by $\Omega^{\perp}$. If we define an orthogonal projection operator P, the matrix $\mathbf{X}$ can be projected onto the linear space of matrices supported by Ω [15]:

$$P_{\Omega}(\mathbf{X})_{(i,j)} = \begin{cases} X_{ij} & if(i,j) \in \Omega \\ 0 & if(i,j) \notin \Omega \end{cases} \tag{4}$$

Now the matrix completion problem in Lagrange form (3) can be written in a nice form:

$$\underset{Z}{minimize}\ \frac{1}{2}||P_{\Omega}(\mathbf{X}) - P_{\Omega}(\mathbf{Z})||_F^2 + \lambda||\mathbf{Z}||_* \tag{5}$$

3.2. Lemma

To solve the optimization problem (5), we first present the following lemma (proof can be found in [9]).

Lemma. If matrix $\mathbf{W}_{m\times n}$ has rank r, then the optimization problem:

$$\underset{Z}{min}\ \frac{1}{2}||\mathbf{W} - \mathbf{Z}||_F^2 + \lambda||\mathbf{Z}||_* \tag{6}$$

has solution $\widehat{\mathbf{Z}} = \mathbf{S}_{\lambda}(\mathbf{W})$, where

$$\begin{aligned} &\mathbf{S}_{\lambda}(\mathbf{W}) = \mathbf{U}\mathbf{D}_{\lambda}\mathbf{V}^T \\ &with\ \mathbf{D}_{\lambda} = diag[(d_1 - \lambda)_+, ..., (d_r - \lambda)_+] \end{aligned} \tag{7}$$

$\mathbf{UDV}^T$ is the Singular Value Decomposition *(SVD)* of $\mathbf{W}$, here $t_+ = max(t, 0)$. The notation$\mathbf{S}_{\lambda}(\mathbf{W})$ refers to *soft-thresholding* [20]

3.3. Soft-Impute Algorithm

Now we begin to introduce Soft-Impute Algorithm. First we rewrite (5) as follows:

$$\begin{aligned} &\underset{Z}{minimize}\ \frac{1}{2}||\mathbf{X} - \mathbf{Z}||_F^2 + \lambda||\mathbf{Z}||_* \\ &= \underset{Z}{minimize}\ \frac{1}{2}||P_{\Omega}(\mathbf{X}) - [\mathbf{Z} - P_{\Omega^{\perp}}(\mathbf{Z})]||_F^2 + \lambda||\mathbf{Z}||_* \\ &= \underset{Z}{minimize}\ \frac{1}{2}||[P_{\Omega}(\mathbf{X}) + P_{\Omega}^{\perp}(\mathbf{Z})] - \mathbf{Z}||_F^2 + \lambda||\mathbf{Z}||_* \end{aligned} \tag{8}$$

By Lemma in part 3.2, the optimal solution of optimization problem (8) can be solved by iteratively updating $\mathbf{Z}$ using

$$\mathbf{Z} \longleftarrow \mathbf{S}_{\lambda}(P_{\Omega}(\mathbf{X}) + P_{\Omega^{\perp}}(\mathbf{Z})) \tag{9}$$

With an arbitrary initialization.

As for the parameter tunning, we propose a cross-validation-like strategy to select the optimal one. The idea is as follows: Ω is the index of observed entries of $\mathbf{X}$. Firstly we randomly introduce an additional 5% artificial deletions in Ω, by deleting the original observed ones to get a test dataset. Then we solve (5) on a grid of λ values on the test dataset. We start from a large λ_{max}, which equals to the second largest singular value of matrix $P_{\Omega}(\mathbf{X})$. We set the maximum rank of $\mathbf{Z}$, denoted as $rank_{max}(\mathbf{Z})$, equal to min(m, n). If $rank(\mathbf{Z}) < rank_{max}(\mathbf{Z})$, we continue solving (5), and reduce λ by a factor $\eta = 0.9$, until $rank(\mathbf{Z}) \geq rank_{max}(\mathbf{Z})$. Finally, to select the optimal parameter, we evaluate the prediction error between the actual data and the predicting data on a grid of λ on the test dataset. Here the *Person correlation* and the *normalized root mean squared error (NRMSE)* (10) are used as the evaluation criterion. We can choose the parameter λ^*, which minimizes the prediction error (the highest Person correlation or lowest NRMSE).

$$NRMSE = \sqrt{\frac{mean[(ij_{answer} - ij_{guess})^2]}{variance[ij_{answer}]}} \tag{10}$$

Now we have the algorithm:

Algorithm: Soft-Impute

1. Initialize $\mathbf{Z}^{old} = 0$
2. For λ_i in the grid of λ, do from λ_{max}
   (a) Repeat:
      i. Compute $\mathbf{Z}^{new} \leftarrow \mathbf{S}_{\lambda_i}(P_\Omega(\mathbf{X}) + P_{\Omega^\perp}(\mathbf{Z}^{old}))$
      ii. Define the energy function: $f_{\lambda_i}(\mathbf{Z}) = \frac{1}{2}\|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{Z})\|_F^2 + \lambda\|\mathbf{Z}\|_*$
      if $\frac{|f_{\lambda_i}(\mathbf{Z}^{new}) - f_{\lambda_i}(\mathbf{Z}^{old})|}{f_{\lambda_i}(\mathbf{Z}^{old})} < \varepsilon$ exit.
      iii. $\mathbf{Z}^{old} \leftarrow \mathbf{Z}^{new}$
   (b). Assign $\widehat{\mathbf{Z}}_{\lambda_i} \leftarrow \mathbf{Z}^{new}$

   Then $\lambda_{i+1} = \lambda_i * 0.9$ for λ_{i+1} repeat (a) to (b) until : $rank(\mathbf{Z}) \geq rank_{max}(\mathbf{Z})$
3. Output the solutions: $\widehat{\mathbf{Z}}_{\lambda_1}, ..., \widehat{\mathbf{Z}}_{\lambda_i}, \widehat{\mathbf{Z}}_{\lambda_{i+1}}, ..., \widehat{\mathbf{Z}}_{\lambda_{max}}$
4. Choose the optimal solution: $\widehat{\mathbf{Z}}_{\lambda_{optimal}}$

4. Results

4.1. Assessing the accuracy of quantitative imputation

We applied the Soft-Impute to three genetic interaction datasets and one synthetic dataset. Three genetic interaction datasets investigated here are ESP-EMAP, CHR-EMAP and S. pombe global genetic interaction map.

To assess the effectiveness of imputation techniques for genetic interaction datasets (EMAP), an approach we adopt is to artificially introduce additional missing values on base of an existing incomplete EMAP matrix. We randomly delete the original observed data in EMAP or synthetic dataset, to get a new matrix with artificial missing entries. This process is repeated multiple times so that we get a series of test matrixes for every original EMAP or synthetic data matrix. After applying different imputation methods on these test matrixes, we can evaluate the prediction error between the actual data and the predicting data. Then we get one imputation accuracy distribution for each EMAP dataset in terms of each imputation method.

For ESP, CHR dataset and the synthetic dataset we repeat the artificial deletions for twenty times, resulting in twenty test matries respectively. For the S. pombe set, which containing about 1.6 million pairwise data, is too computational expensive, we repeat 15 times only.

There are several existing methods to impute missing values that have been used in synthetic data. These algorithms (see Table 1) include three local methods: k-Nearest Neighbors imputation (kNN) [21], Local Least Squares imputation (LLS) [22], Iterated Local Least Square imputation (iLLS) [23]; and one global method: Bayesian Principal Component Analysis imputation (BPCA) [24]. The "local" here represent those algoriths impute missing values through local information around the missing value.

Figure. 1a-b show when 10% rate of artificial missing values is generated for 20 times in synthetic dataset, Soft-Impute algorithm performs best on all test matrixes no matter Person correlation or NRMSE evaluated. BPCA algorithm could not get one stable result, which has been demonstrated in [24] that when genes have dominant local similarity structures, BPCA doesn't work a good performance. Our test matrixes induced from synthetic dataset have high local correlation which refers to a gene module, but BPCA algorithm is limited by this characteristic. In order to demonstrate the performance of Soft-Impute algorithm on dataset with different rates of missing values. Artificial missing rate different from 1% to 37% is used. Figure. 1c-d gives the imputing results of this gradient missing rate in 20 synthetic data test matrixes. To demonstrate whether such accuracy performance differences in synthetic dataset are statistically significant, we presented the t-test for every two distributions and derived the statistical significance (P-value) (see Figure. 2 a-b). The running time showed in Table 1 are computed when these imputation methods applied on 20 synthetic data test matrixes with 10% artificial deletion values and gradient artificial missing value rate.

Table 1: Running Time of Imputation Methods on Synthetic Datasets

Imputation Methods	Implementation	CPU running time(/)
Soft-Impute	Matlab	616.0s/650.1s
Local Least Squares (LLS)	Matlab	732.1s/4407.8s
Iterated Local Least Squares (ILLS)	Matlab	2367.2s/19987.3s
Bayesian Principle Component (BPCA)	Matlab	10198.0s/11037.8s
k Nerest Neighbor (kNN)	R	159.9s170.2s

[1] Running time (10% missing values/gradient missing rate from 1% to 37%)

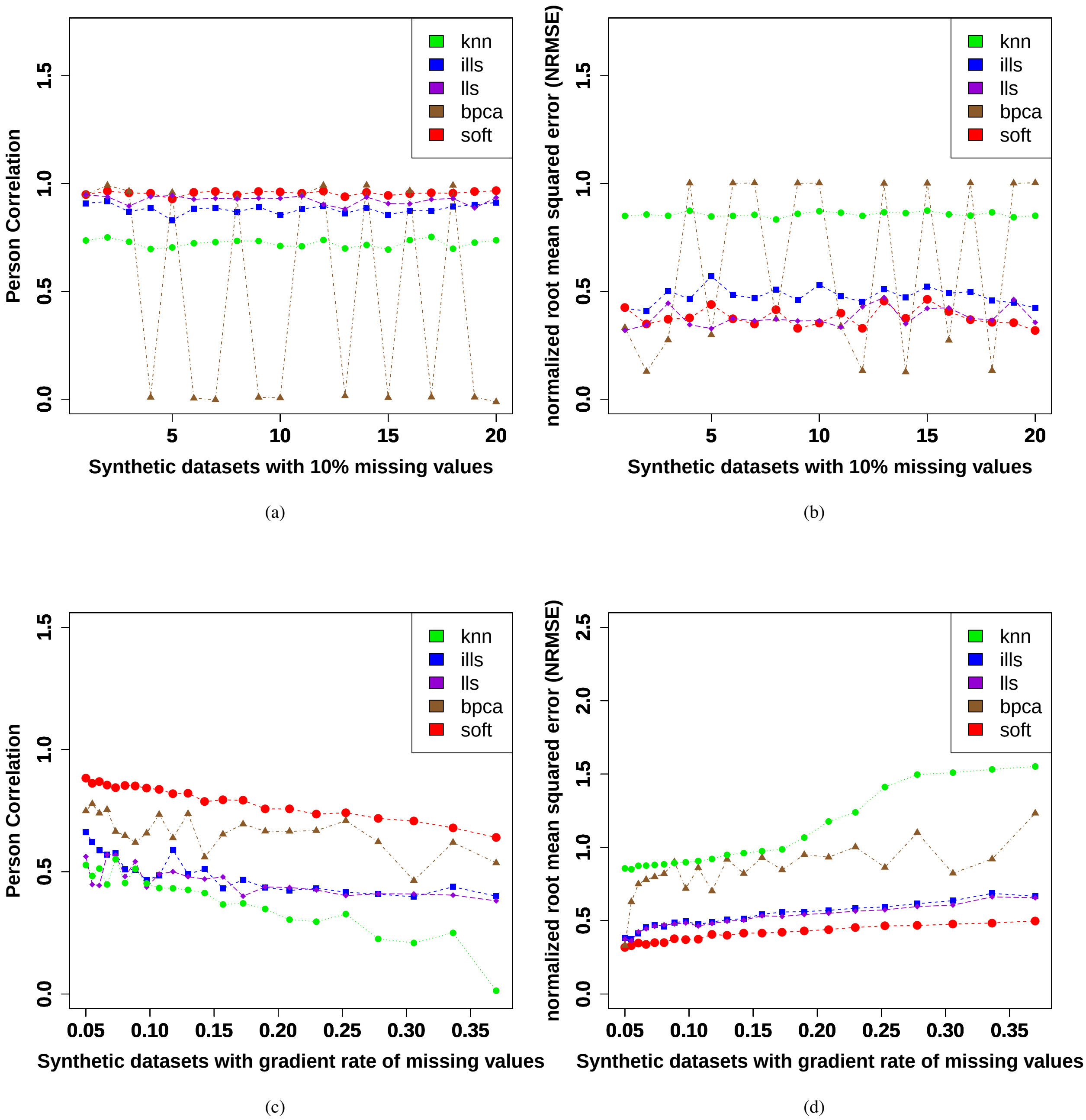

Figure 1: Capability of the imputation methods to reproduce the original measurements in the synthetic datasets. a-b showed the imputation accuracy (person correlation and NRMSE) on synthetic datasets with 10% values deleted. c-d showed the results on the same original synthetic dataset with gradient missing rates, the x axis represents the missing rate.

Imputation accuracies in terms of person correlation and NRMSE of the different imputation methods on 20 ESP-EMAP data test matrixes with artificial 10% deletion values are shown in Figure. 3a-b. Within each matrix, the best imputation accuracies are obtained by Soft-Imputation, and BPCA algorithm is also not stable to get one optimal result. The second best performing method are LLS and ILLS, which cost more time (see Table 2). To demonstrate the imputation results across different missing rate on ESP-EMAP, we also construct the gradient artificial missing rate from 1% to 25%, by introducing artificial missing entries in the observed subset (see Figure. 3c-d). For each method, the imputation accuracy decreases with the increasing missing value rate, meanwhile the Soft-Impute method almost gets a better performance for different missing rate. Because of the high rate of missing values in the original datasets of CHR-EMAP and S. pombe, we didn't evaluate the imputing accuracy with gradient missing rates. For CHR-EMAP and S. Pombe dataset we introduce 5% artificial missing values. The BPCA method depends heavily on the properties of the dataset being imputed, and therefore it was difficult to find a stable optimal point. This time, in terms of S. Pombe dataset, BPCA could not converge, so there is no result of BPCA.

Running times of the different imputation methods on these three EMAP datasets are presented in Table 2. k Nearest Neighbor imputation is faster than the more advanced imputation methods, but has the worst imputation accuracy (Figure. 3). Soft-Impute algorithm costs a median time but gets the best accuracy among all the datasets. Its surprisingly imputation ability to the EMAP datasets is supported by the results of Figure. 3.

We have demonstrated the process of constructing EMAP and synthetic data test matrixes with random artificial missing values. So for every EMAP dataset (ESP, CHR, Pombe) and synthetic one, there are several distributions of the imputation accuracy in terms of different imputation methods. To demonstrate the imputation accuracy differences on EMAP datasets with different imputation methods are statistically significant, for every EMAP dataset, we derived statistic significance (P-value) of t-test between the accuracy distribution of Soft-Imputation and that of the other methods. This result is showed in Figure. 2. In Figure. 2, X axis represents methods, and Y axis represents the means of imputation accuracy.

Table 2: Running Time of Imputation Methods on ESP Datasets

EMAP Datasets	Imputation Methods				
	Soft-Impute	k-Nearest Neighbors(kNN)	Local Least Squares(LLS)	Iterated Local Least Square(iLLS)	Bayesian Principal Component(BPCA)
ESP with 10% missing values	1537.4s	172.6s	3756.6s	7572.4s	46558.3s
ESP with grad missing values	2006.3s	288.9s	4961.3s	9015.3s	55126.7s
CHR with 5% missing values	31681.3s	1568.5s	38789.3s	52817.9s	668890.7s
Pombe with 5% missing values	138979.3s	2327.8s	82647s	1.3141e+05s	NA

For example, in Figure. 2a-b there is a set of 20 test matrixes with artificial deletions induced from the original synthetic dataset. After imputing on this set with different imputation methods, there are two accuracy distributions: person correlation (Figure. 2a) and NRMSE (Figure. 2b) for every imputation method. In this two figures, the means of accuracy distributions of different imputation methods are plotted in Figure. 2. Where the red histogram represents Soft-Impute method, and blue ones represents the other imputation methods. In the two figures, the p-values after t-test have also been showed.

Figure. 2 shows means of imputation accuracy in terms of Person correlation and NRMSE of different methods on three kinds of EMAP datasets and synthetic dataset. Through the t-test, we can see that the better imputation accuracy of Soft-Impute algorithm than other methods on the three kinds of EMAP datasets and the synthetic dataset is statistically significant.

Algorithms such as kNN, LLS and iLLS focus on the local information around miss values, for example, KNN only depends on the nearest K neighbours for each missing entry. BPCA does not have a good performance, when genes have dominant local similarity structures [24]. However, Soft-Impute method performs SVD algorithm to control the whole rank of the matrix, and chooses one optimal parameter to control the rank. This methodology takes full advantage of the correlation of the genetic interaction profile to predict unknown genetic interactions.

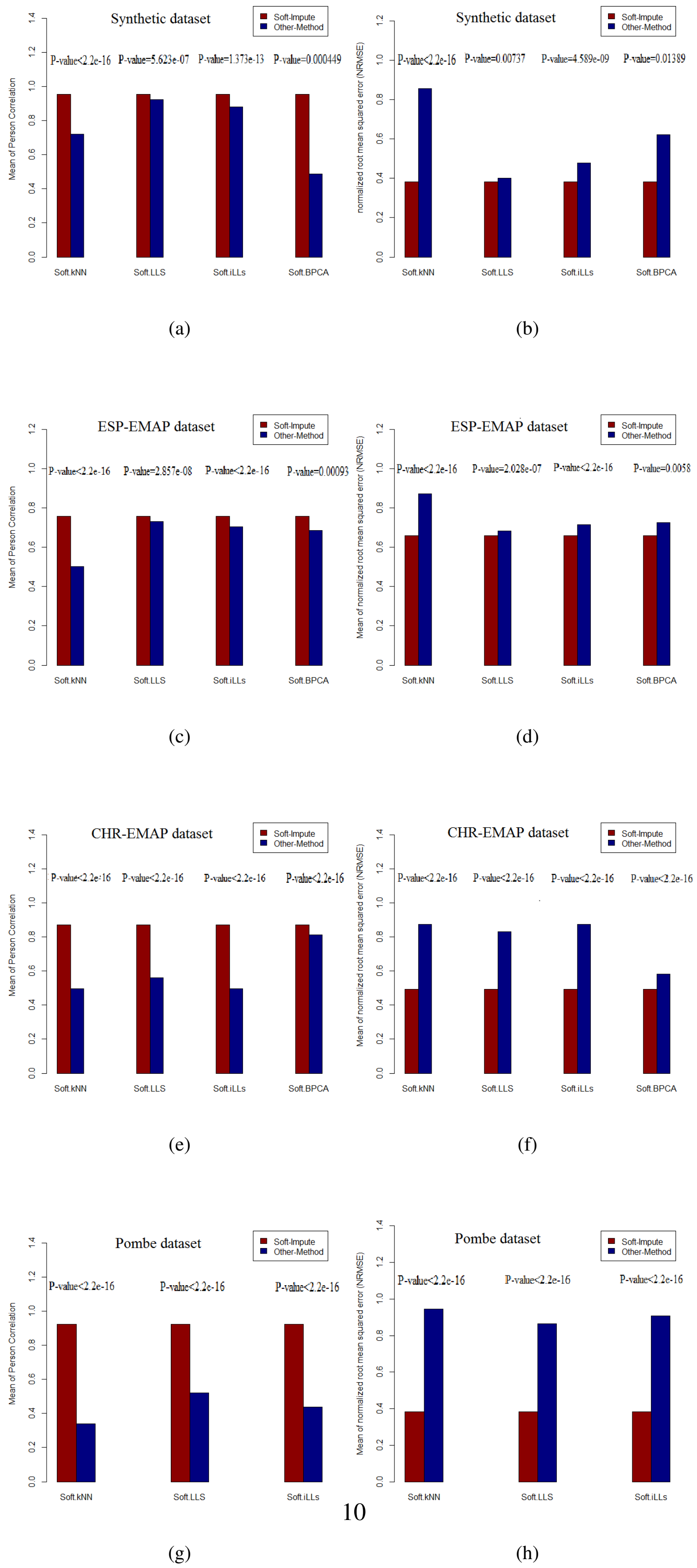

Figure 2: Comparison of Soft-Impute with other methods on means of accuracy distributions.

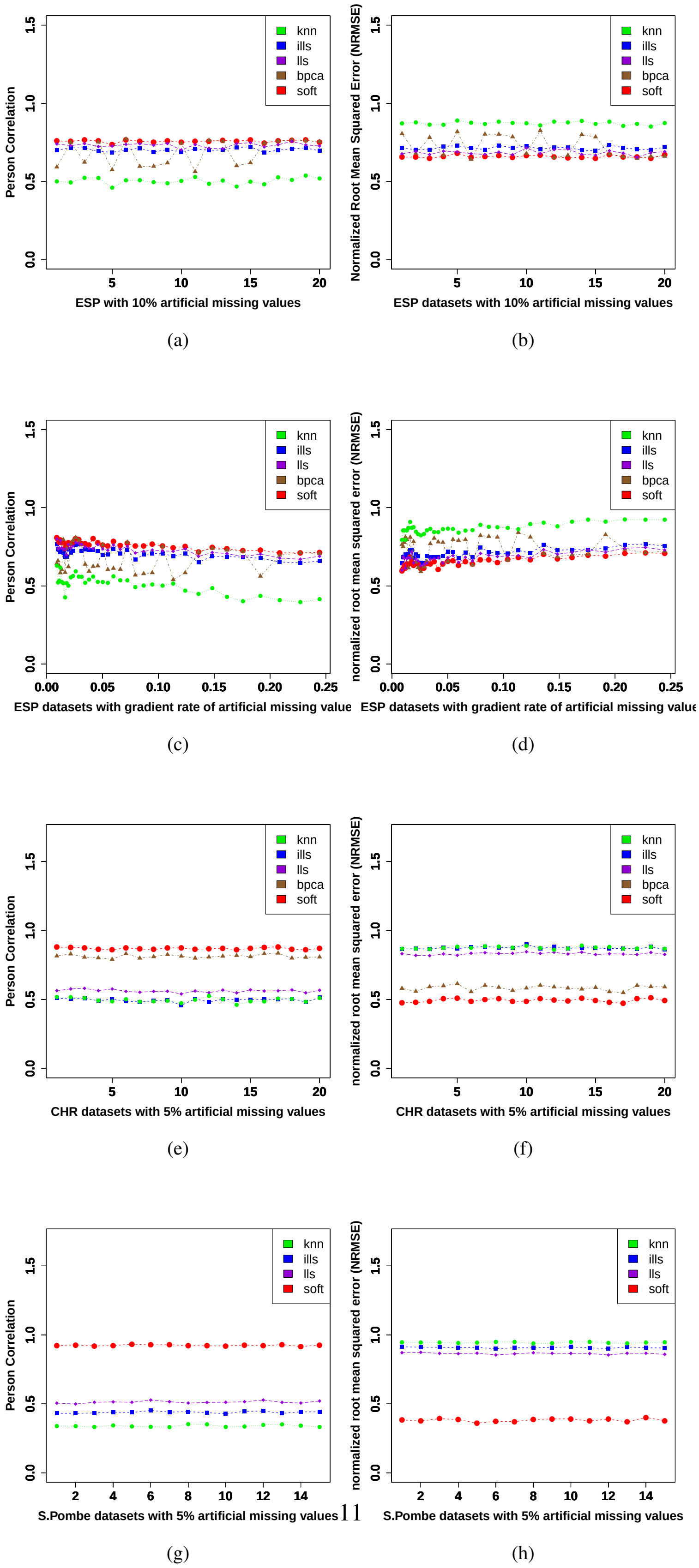

Figure 3: The imputation accuracy of imputation methods on EMAP datasets.

4.2. Agreement with the original clustering results

Another of our motivations to impute in EMAP datasets is to improve downstream analysis. A widely used analysis technique applied to EMAP datasets is average-linkage hierarchical clustering, using R program. In order to assess the impact of different imputation methods on clustering and downstream biological analysis, we compared clustering results on the three kinds of EMAP datasets, after imputation using kinds of methods presented above. Soft-Impute algorithm improved the clustering results better than the other imputation methods. We used the Jaccard index to determine how well the predicted cluster modules correspond to benchmark gene sets (GO terms). The Jaccard index [25] between two sets M_i and B_j is defined as:

$$\frac{\sharp\{M_i \cap B_j\}}{\sharp\{M_i \cup B_j\}} \tag{11}$$

where $\sharp\{A\}$ denotes the number of set A

For module M_i, the Jaccard Index between M_i and the benchmark gene sets is defined as the maximum of Jaccard Index between M_i and any gene set in the benchmark:

$$Jaccard\ Index(M_i, B) = max_j\{JaccardIndex(M_i, B_j)\} \tag{12}$$

Thus, the average Jaccard index of the predicted modules and the benchmark gene sets can be computed:

$$Jaccard\ Index(M, B) = \frac{\sum\limits_{i \in 1,\ldots k} Jaccard\ Index(M_i, B)}{k} \tag{13}$$

The accuracy of clustering result is evaluated by average Jaccard index of the predicted modules and benchmark gene sets. In the ideal situation where the predicted modules perfectly match the benchmark gene sets, the Jaccard index is 1. The larger the Jaccard index, the better the predictions are. Hierarchical clustering algorithm is used to predict the gene clusters in the three kinds of original EMAP gene sets after kinds of imputation. The benchmark ("theoretical") gene sets are GO iterms. The results are presented as Jaccard index, numbers of predicted gene modules, and numbers of predicted gene modules enriched in GO (see Table. 3). The clear differences can be observed in Table. 3 among different imputation methods.

Table 3: Clustering Results

Imputation Methods	Jaccard Index	$\sharp$ Modules	$\sharp$ Enriched Benchmark set[@]
Soft-Impute	0.118	41	35
LLS	0.102	14	13
iLLS	0.086	15	13
BPCA	0.093	17	15
kNN	0.097	26	22

A : ESP-EMAP datasets

Imputation Methods	Jaccard Index	♯ Modules	♯ Enriched Benchmark set[@]
Soft-Impute	0.075	74	70
LLS	0.063	24	36
iLLS	0.062	55	53
BPCA	0.062	40	39
kNN	0.073	54	54

B : CHR-EMAP datasets

Imputation Methods	Jaccard Index	♯ Modules	♯ Enriched Benchmark set[@]
Soft-Impute	0.029	188	143
LLS	0.024	67	55
iLLS	0.023	61	53
kNN	0.022	60	51

C : S. Pombe-EMAP datasets

@: hyper-geometric test applied to test the enrichment of gene sets. Significance level :FDR<=0.05. $\sharp Modules$: the number of modules predicted by hierarchical clustering after EMAP datasets imputed by different imputation methods. $\sharp EnrichedBenchmarkset$: the number of modules predicted by hierarchical clustering enriched in the GO-slim iterms

Table. 3 (A-B) present the results of S. Cerevisiae EMAP datasets. GO-slim iterm of S. Cerevisiae was downloaded from SGD. Table. 3 (C) presents the result of S. Pombe EMAP dataset, and GO-slim iterm of S. Pombe was downloaded from the home page of Prof. Krogan (http://kroganlab.ucsf.edu/).

The EMAP datasets clustering results after imputed by kinds of imputation methods are finally compared by measuring their consistency with known gene modules (GO iterms). For ESP-EMAP dataset [11], there are many functionally homogeneous subtrees of the quantitative EMAP score matrix. The author [11] presented several interaction-dense clusters. For CHR-EMAP dataset and S. Pombe dataset, there are many previously detected modules [12] [13]. We also compared the predict modules got by clustering EMAP datasets after imputation by kinds of methods with the published gene modules. The results can be found in Table. 4. We compared the number of genes predicted by hierarchical clustering that are in the published modules, and the number of published modules in which gene are predicted by hierarchical clustering.

Table 4: Clustering Results

Imputation Methods	♯ Genes in published modules	♯ Published modules found by h-cluster	♯ Gene number mean of published modules
Soft-Impute	63	6	10.5
LLS	37	4	9.25
iLLS	37	4	9.25
BPCA	39	4	9.75
kNN	6	1	6

A : ESP-EMAP datasets

Imputation Methods	♯ Genes in published modules	♯ Published modules found by h-cluster	♯ Gene number mean of published modules
Soft-Impute	36	6	6
LLS	30	6	5
iLLS	30	6	5
BPCA	30	6	5
kNN	2	1	2

B : CHR-EMAP datasets

Imputation Methods	♯ Genes in published modules	♯ Published modules found by h-cluster	♯ Gene number mean of published modules
Soft-Impute	68	14	4.9
LLS	48	9	5.3
iLLS	43	12	3.6
kNN	46	13	3.5

C : S. Pombe-EMAP datasets

Table. 4 presents that our method applying imputation before hierarchical clustering (using average-linkage) are more informative. Specially, in S. Pombe dataset, [13] has demonstrated that they found two previously uncharacterized genes: SPAC1610.01 and SPAC18G6.13 which were clustered in mRNA splicing module. The clustering results after Soft-Impute imputation could find this module containing these two genes and other genes involved in mRNA splicing, while LLS, iLLS or kNN can not find these genes.

These results have demonstrated the ability of Soft-Impute algorithm to improve the downstream data analysis. Soft-Impute Algorithm takes advantage of the correlation of genetic interaction profile to predict unknown genetic interactions. The $\mathbf{Z}$ matrix is the integrated matrix that has been imputed by Soft-Impute imputation and its rank is limited during the imputation. Mathematically, the Soft-Impute procedure eliminates those small eigenvalues and reserve big eigenvalues, equivalently, such a procedure tend to clear up data to enhance the strong correlation structure among genes in the data matrix. In other words, the procedure of Soft-Impute algorithm achieves the reorganization and refining of the original dataset while imputes the missing entries. So this methodology could better improve the downstream clustering effectiveness while predicting the missing entries than the other imputation methods do.

5. Conclusion

In this article, we have introduced a method named Soft-Impute to impute missing values of EMAP datasets. Soft-Impute method uses the correlation among genes to impute the missing values. This method adopts an efficient algorithm to solve imputation problem and guaranteed its convergence [9]. It develops one soft threshold to SVD algorithm, which can be selected an optimal one by choosing the regularization parameter λ. This methodology was proposed by Hastie [9], and has been used in image recovery and eQTL study, but it is the first time to be introduced in genetic interaction data imputation.

We have compared Soft-Impute method with four other popular imputation methods for genetic interaction data, on one synthetic dataset and tree kinds of EMAP datasets. Firstly, the given datasets were imputed and evaluated the imputation accuracy, and then clustered by hierarchical clustering. Finally we compared the clustering results against GO annotations, and the published literature annotations. We demonstrated that Soft-Impute method achieved a better performance in imputation accuracy and improved downstream data analysis than other existing methods.

As far as we know, this paper is the first attempt to introduce Soft-Impute algorithm into imputing missing values in genetic interaction datasets.This algorithm is appropriate for datasets where there are modules in which entries are with high correlation. So it can be used widely in many kinds of fields with such characteristics to realize the imputation of missing entries.

The imputation of missing values is the first step of data analysis, and it has very important impact to the downstream analysis. Soft-Impute method could improve the performance of downstream data analysis and promote the further exploration of genetic interaction network.

6. Acknowledgment

This work is supported by the National Natural Science Foundation of China (31171262, 11021463), the National Key Basic Research Project of China (2009CB918503), and partly supported by MOST international collaborative project (2011DFA31860).

[1] R. A. Fisher, Xv.łthe correlation between relatives on the supposition of mendelian inheritance., Transactions of the Royal Society of Edinburgh 52 (02) (1919) 399–433.
[2] C. Boone, H. Bussey, B. J. Andrews, Exploring genetic interactions and networks with yeast, Nature Reviews Genetics 8 (6) (2007) 437–449.
[3] S. R. Collins, M. Schuldiner, N. J. Krogan, J. S. Weissman, A strategy for extracting and analyzing large-scale quantitative epistatic interaction data, Genome biology 7 (7) (2006) R63.
[4] C. Ryan, D. Greene, G. Cagney, P. Cunningham, Missing value imputation for epistatic maps, BMC bioinformatics 11 (1) (2010) 197.
[5] C. Ryan, G. Cagney, N. Krogan, P. Cunningham, D. Greene, Imputing and predicting quantitative genetic interactions in epistatic maps, in: Network Biology, Springer, 2011, pp. 353–361.
[6] O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, The mystery of missing heritability: genetic interactions create phantom heritability, Proceedings of the National Academy of Sciences 109 (4) (2012) 1193–1198.
[7] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, et al., The biogrid interaction database: 2011 update, Nucleic acids research 39 (suppl 1) (2011) D698–D704.
[8] A. Roguev, D. Talbot, G. L. Negri, M. Shales, G. Cagney, S. Bandyopadhyay, B. Panning, N. J. Krogan, Quantitative genetic-interaction mapping in mammalian cells, Nature methods.
[9] R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, The Journal of Machine Learning Research 99 (2010) 2287–2322.
[10] C. Yang, L. Wang, S. Zhang, H. Zhao, Accounting for non-genetic factors by low-rank representation and sparse regression for eqtl mapping, Bioinformatics 29 (8) (2013) 1026–1034.
[11] M. Schuldiner, S. R. Collins, N. J. Thompson, V. Denic, A. Bhamidipati, T. Punna, J. Ihmels, B. Andrews, C. Boone, J. F. Greenblatt, et al., Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile, Cell 123 (3) (2005) 507–519.
[12] S. R. Collins, K. M. Miller, N. L. Maas, A. Roguev, J. Fillingham, C. S. Chu, M. Schuldiner, M. Gebbia, J. Recht, M. Shales, et al., Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map, Nature 446 (7137) (2007) 806–810.
[13] C. J. Ryan, A. Roguev, K. Patrick, J. Xu, H. Jahari, Z. Tong, P. Beltrao, M. Shales, H. Qu, S. R. Collins, et al., Hierarchical modularity and the evolution of genetic interactomes across species, Molecular cell 46 (5) (2012) 691–704.
[14] N. Srebro, N. Alon, T. S. Jaakkola, Generalization error bounds for collaborative prediction with low-rank matrices, in: Advances In Neural Information Processing Systems, 2004, pp. 1321–1328.
[15] J.-F. Cai, E. J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM Journal on Optimization 20 (4) (2010) 1956–1982.
[16] E. J. Candès, T. Tao, The power of convex relaxation: Near-optimal matrix completion, Information Theory, IEEE Transactions on 56 (5) (2010) 2053–2080.
[17] R. H. Keshavan, A. Montanari, S. Oh, Matrix completion from a few entries, Information Theory, IEEE Transactions on 56 (6) (2010) 2980–2998.
[18] N. Srebro, T. Jaakkola, et al., Weighted low-rank approximations, in: ICML, Vol. 3, 2003, pp. 720–727.
[19] M. Fazel, Matrix rank minimization with applications, Ph.D. thesis, PhD thesis, Stanford University (2002).
[20] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, D. Picard, Wavelet shrinkage: asymptopia?, Journal of the Royal Statistical Society. Series B (Methodological) (1995) 301–369.
[21] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman, Missing value estimation methods for dna microarrays, Bioinformatics 17 (6) (2001) 520–525.
[22] H. Kim, G. H. Golub, H. Park, Missing value estimation for dna microarray gene expression data: local least squares imputation, Bioinformatics 21 (2) (2005) 187–198.
[23] Z. Cai, M. Heydari, G. Lin, Microarray missing value imputation by iterated local least squares., in: APBC, 2006, pp. 159–168.
[24] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, S. Ishii, A bayesian missing value estimation method for gene expression profile data, Bioinformatics 19 (16) (2003) 2088–2096.
[25] J. Song, M. Singh, How and when should interactome-derived clusters be used to predict functional modules and protein function?, Bioinformatics 25 (23) (2009) 3143–3150.