The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 16):493
DOI 10.1186/s12859-016-1308-y

BMC Bioinformatics

RESEARCH Open Access

Spectral consensus strategy for accurate reconstruction of large biological networks

Séverine Affeldt[1†], Nataliya Sokolovska[1,2,3†], Edi Prifti[1] and Jean-Daniel Zucker[1,2,4*]

From The 10th International Workshop on Machine Learing in Systems Biology (MLSB)
Den Haag, The Netherlands. 3-4 September 2016

Abstract

Background: The last decades witnessed an explosion of large-scale biological datasets whose analyses require the continuous development of innovative algorithms. Many of these high-dimensional datasets are related to large biological networks with few or no experimentally proven interactions. A striking example lies in the recent gut bacterial studies that provided researchers with a plethora of information sources. Despite a deeper knowledge of microbiome composition, inferring bacterial interactions remains a critical step that encounters significant issues, due in particular to high-dimensional settings, unknown gut bacterial taxa and unavoidable noise in sparse datasets. Such data type make any a priori choice of a learning method particularly difficult and urge the need for the development of new scalable approaches.

Results: We propose a consensus method based on spectral decomposition, named *Spectral Consensus Strategy*, to reconstruct large networks from high-dimensional datasets. This novel unsupervised approach can be applied to a broad range of biological networks and the associated spectral framework provides scalability to diverse reconstruction methods. The results obtained on benchmark datasets demonstrate the interest of our approach for high-dimensional cases. As a suitable example, we considered the human gut microbiome co-presence network. For this application, our method successfully retrieves biologically relevant relationships and gives new insights into the topology of this complex ecosystem.

Conclusions: The *Spectral Consensus Strategy* improves prediction precision and allows scalability of various reconstruction methods to large networks. The integration of multiple reconstruction algorithms turns our approach into a robust learning method. All together, this strategy increases the confidence of predicted interactions from high-dimensional datasets without demanding computations.

Keywords: Network reconstruction, Community-based method, Spectral theory, High-dimensional data, Microbiota

Background

Discovering complex interactions is a long-standing problem which led over the past years to the development of many network reconstruction methods that exhibit competitive results on various types of data. As successfully demonstrated, networks are invaluable tools to comprehensively relate biological variables [1–3] and possibly gain insights into their direct causal relationships [4]. Interestingly, recent studies have shown that the available approaches would not generally perform optimally across all dataset types and the integration of diverse inference methods can provide an improved robust performance [5–8]. However, several well-known and widely used algorithms cannot directly process high-dimensional data or actually perform better on small networks. Bringing these methods within a lower dimensional space would enable researchers to fully benefit from their strengths under high-dimensional settings, and more

*Correspondence: jean-daniel.zucker@ird.fr
†Equal contributions
[1]Integromics, Institute of Cardiometabolism and Nutrition, ICAN, Assistance Publique Hôpitaux de Paris, Pitié-Salpêtrière Hospital, Paris 75013, France
[2]Sorbonne Universités, UPMC University Paris 6, UMR S U1166 NutriOmics Team, Paris 75013, France
Full list of author information is available at the end of the article

interestingly, to integrate their outcome in community-based predictions.

We propose a *consensus* approach, named *Spectral Consensus Strategy* (SCS), to reconstruct complex biological networks from high-dimensional datasets. This method provides scalability to various reconstruction methods and can be applied to a broad range of complex biological networks. Our approach unfolds in three parts. First, it relies on a spectral framework to identify sets of significantly related variables. Specifically, the subset selection uses the magnitude of the normalized Laplacian eigenvector elements. These subsets are then considered in a second phase for multiple parallel *local* reconstructions from which global effects are inferred. By enabling each reconstruction method to locally *avoid* high-dimensional settings, this second phase improves individual prediction accuracy and scalability. In the last phase, the individual reconstructions that benefited from the spectral embedding are integrated in a consensus network.

All together, this strategy provides robust and accurate reconstructions from high-dimensional observational data for which no suitable learning approach is known beforehand, as for instance frequently encountered in metagenomics. To our knowledge, our contribution is the first attempt to introduce a consensus network reconstruction approach based on a spectral framework.

Network reconstruction background

Generally speaking, network learning algorithms can be divided into two categories: *constraint-based* and *score-based* approaches. The constraint-based methods ascertain (conditional) independence relationships from statistical tests [9, 10] to learn structural constraints in causal graphs. These approaches are highly efficient on sparse networks and are guaranteed to learn the Markov equivalent class of the underlying graphical model if the *exact* list of conditional independence relationships is given. However, constraint-based methods have also proved to be very sensitive to sampling noise from finite datasets. Alternatively, score-based methods identify the model that best fits the data through the maximization of a score function over the space of (ideally all) possible Bayesian networks [11, 12]. To learn the networks in reasonable time, the search procedure usually follows a heuristic algorithm that identifies a local optimum. More recently, several *mutual information-based* approaches have been proposed to infer direct relationships from noisy observational datasets containing few samples [1, 2]. Nevertheless, as demonstrated by the growing number of *hybrid* approaches [4, 13–15], the wide range of high-dimensional data is still challenging state-of-the-art methods, both in terms of accuracy, or time and memory consumption.

Spectral methods background

Spectral theory has provided a number of approaches to uncover dataset structure. A well-known result is the ability to optimally bi-partition a graph based on the second eigenvector of the normalized Laplacian matrix, also known as *algebraic connectivity* or *Fiedler vector* [16, 17]. Following this idea, recursive *two-way* cut methods [18–20] that rely solely on the second eigenvector, and *k-way* cut approaches [21–26] that are based on truncated eigenvector basis, have been successfully applied to dimensionality reduction or clustering problems. Specifically, the truncated eigenvector basis provides a new representation that amplifies the similarity between closely related variables while reducing the affinity of unrelated variables [26–29]. Many biological systems are usually composed of overlapping sub-units that involve functionally related features, such as found in metabolic or gene regulatory networks. Hence, learning large biological networks from multiple local reconstructions appears to be a reasonable procedure as much as it follows the natural dataset structure. Spectral methods hold great potential for guiding learning algorithms that perform better on small graphs towards improving inference of large networks.

Consensus reconstruction approaches

The idea of consensus or *ensemble* learning is recently gaining interest in the field. An example is given in [30] where the yeast metabolic network was reconstructed based on a complex consensus procedure that involved a number of statistical methods and an important amount of prior knowledge. As previously demonstrated [31], consensus approaches can be efficiently exploited to reconstruct Bayesian networks and provide robust models from biological data. A consensus method that mainly rely on significance tests is proposed in [32] to learn dependencies between gene regulatory factors in the human frontal lobe, resulting in a high-confidence model. The community structure in complex networks can also be revealed by consensus clustering as reported in [33], where a stable partitioning approach based on several stochastic method results is proposed. Marbach et al. [5] motivates the development of consensus methods by demonstrating the benefits of combining complementary inference approaches. Specifically, they have evaluated the performance of diverse learning algorithms and shown that their combination performs robustly across various datasets while providing as good or better results than individual methods.

The complex gut microbiome system

The human gut hosts a high density of commensal bacteria whose collective genome, also known as *metagenome*, exceeds more than a hundred times the size of the human

genome [34]. This rich ecosystem provides the host with vital functions that affect nutritional efficiency and overall health [35, 36]. Over the past few years, the role of gut microbiota in human health has received unprecedented attention [37]. In particular, several chronic diseases such as obesity [38, 39], inflammatory bowel disease [40, 41], liver cirrhosis [42, 43], type-I [44], and type-II diabetes [45, 46] have been associated with gut microbiota. For a long time, the composition of human gut microbial ecosystem was unknown, especially due to the large number of non-cultivable species. The recent availability of metagenomic data along with different binning techniques allows now to obtain a better picture of the taxonomical groups that inhabit the gut microbiome [47]. These species are organized in complex ecological networks and can be involved in different types of interactions such as competition or mutualism [48]. Yet, mapping these relationships with high confidence remains a complicated task for multiple reasons. First of all, as many species are usually absent from one sample to another, metagenomic datasets are very sparse. This sparsity adds on technical artifacts inherent to the obligate multi-step data processing. Hence, metagenomic data are challenging available reconstruction methods, which may individually yield different topologies for the same set of observations.

Methods

We propose a simple yet highly efficient method called *Spectral Consensus Strategy* (SCS) that simultaneously embeds multiple discovery algorithms within a spectral framework for the reconstruction of large graphical model. The strength of the SCS method hinges on two key points that are (i) the accuracy improvement of each individual learning algorithm and (ii) the combination of predictions from complementary reconstruction methods. Specifically, sets of *path-related* variables are first identified based on the magnitude of the graph Laplacian eigenvector elements (Fig. 1,a), then multiple parallel local reconstructions are performed using different learning methods (Fig. 1,b) and lastly a consensus network is built on the previous multiple outcomes (Fig. 1,c).

In the following, we provide theoretical support to the uncovering of connected variable subsets from the first phase of the SCS approach (SCS-*spectral* step, Fig. 1,a). In particular, we demonstrate that subsets of *path-related vertices* can be directly retrieved from the magnitude and sign of individual eigenvector elements. These subsets, which correspond to possibly overlapping dense subgraphs, are given as input to the second phase of the SCS approach (SCS-*learn* step, Fig. 1,b). We finally detail the whole *Spectral Consensus Strategy*.

Normalized Laplacian eigenvectors

We consider the *random-walk* normalized Laplacian matrix L_{rw} as it entails the random walk dynamics from one vertex to another in the corresponding graph $\mathcal{G}$. This matrix is defined as $L_{rw} = I - D^{-1}W$, where I is the identity matrix, $W = (w_{ij})$ is a weight matrix over all pairs of variables and D the diagonal degree matrix with $d_{ii} = \sum_j w_{ij}$.

Community membership indicators

As already established [49], the null eigenvalues of the graph Laplacian matrices are associated with the number of *connected components*. A subset of vertices $A_k \subset V$ is a connected component if (i) all intermediate points that lie on a path between two vertices of A_k also belong to A_k and (ii) there is no connection between the vertices of A_k and its complementary subset $\overline{A_k}$ (Additional file 1: Proposition 1). Interestingly, for the case of finding $k > 2$ clusters, the first k eigenvectors of the normalized Laplacian matrix L_{rw} minimize the *normalized cut* (*NCut*) criterion of the relaxed problem [18, 50],

$$Ncut(A_1, \ldots, A_k) = \frac{1}{2}\sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{vol(A_i)} \quad (1)$$

where $W(A,B) = \sum_{\substack{i\in A \\ j\in B}} \omega_{ij}$, and $vol(A_i) = \sum_{j\in A_i} d_j$.

In a nutshell, the solution of the relaxed *Ncut* minimization problem consists of the orthonormal matrix $H \in \mathbb{R}^{p\times k}$ whose columns are the first k eigenvectors of the normalized Laplacian eigenvector matrix U, associated with the first k smallest eigenvalues.

When the between-cluster similarity is exactly 0, these eigenvectors are the indicator vectors $\{h_j\}_{j\in[1,k]}$ ($h_j \in \mathbb{R}^p$ and $h_j^i = 1$ if $x_i \in A_j$, otherwise 0) of the k connected components [50]. In practice, the distribution of the data points in distinct clusters is hardly encountered, and one should expect the between-cluster similarity to be greater than 0. Yet, under *nearly ideal* conditions the eigenvectors are still close to the indicator vectors, and the elements magnitude and sign of each eigenvector contain information on vertices membership *strength* [18, 50, 51].

Path-related vertices subsets

Beyond the membership indication, the Laplacian matrix eigenvector elements also convey path-relationship information. In the following we assume that v_k is the k-th eigenvector of the normalized Laplacian matrix associated with the connected component A_k. Under ideal conditions, $x_i \in A_k \Rightarrow v_k(i) = 1$, otherwise $v_k(i) = 0$ [50]. In addition, we demonstrate that similar elements of a given eigenvector ($|v_k(i) - v_k(j)| = 0$) indicate path-related variables (x_j is path connected with x_i) based on the Rayleigh quotient [52] (Additional file 1: Proposition 2).

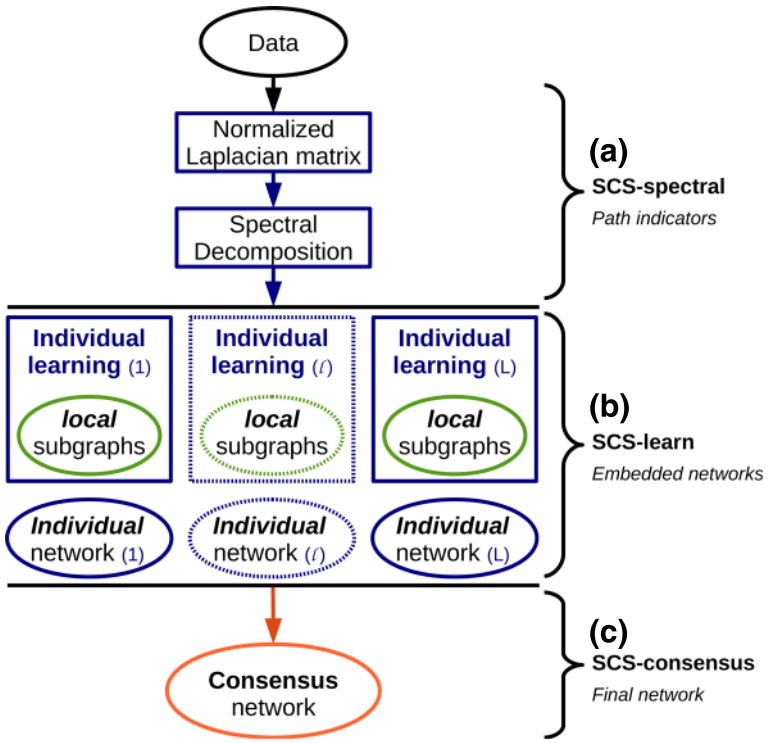

Fig. 1 Overview of the Spectral Consensus Strategy (SCS). The SCS method unfolds in three parts. **a** The SCS-*spectral* phase identifies sets of *path-related* variables based on the magnitude of the graph Laplacian eigenvector elements. **b** The SCS-*learn* phase performs multiple parallel *local* reconstructions using different learning methods. **c** The SCS-*consensus* phase provides a consensus network built on the individual outcomes from the SCS-*learn* step

For the case of a *connected* graph $\mathcal{G}$ (i.e. there is a path between any pair of variables in $\mathcal{G}$) Fiedler's Nodal Domain theorem (Additional file 1: Theorem 1) indicates that while x_i and x_j belong to different clusters A and B, $|v_k(i) - v_k(j)| < \varepsilon$ can be found. However, if there exists a subset of vertices S at a distance less than a *step* $\rho \geq 2$ from A that separates A and B, then v_k is such that [53]

$$\begin{cases} \text{if } i \in A, \text{ then } v_k(i) = 1, \\ \text{if } i \in B, \text{ then } v_k(i) = -1, \\ \text{if } i \in S, \text{ then } -1 + 2/\rho \leq v_k(i) \leq 1 - 2/\rho, \\ \text{if } i, j \text{ are adjacent then } |v_k(i) - v_k(j)| \leq 2/\rho. \end{cases}$$

Taking $\rho = 2$ we obtain the case which is commonly used for separators. Hence, $|v_k(i) - v_k(j)|$ is a measure of the distance between the vertices i and j reflecting the *cluster assumption* which stipulates that close data points are expected to lie within the same cluster (Additional file 1: Proposition 3).

In summary, under ideal conditions, the first k eigenvectors of the normalized Laplacian matrix provide indicator vectors of the k connected components. In practice, the magnitude and sign of the eigenvector elements contain information on vertex membership *strength* to the corresponding component (Additional file 1: Proposition 1). Furthermore, *path-connected* variables have similar eigenvector elements (Additional file 1: Proposition 2), that are distinct from the element of vertices belonging to a different component (Additional file 1: Proposition 3). Thus, subsets of nodes that correspond to large positive or negative eigenvector elements (retrieved in the SCS-*spectral* step) correspond to dense subgraphs (to be reconstructed in the SCS-*learn* step). These subgraphs associated to large eigenvector elements can be redundantly found in the first eigenvectors [54]. However, higher eigenvectors can also be used to identify different subsets of connected nodes, as observed in the context of anomalous graph detection [55].

The spectral consensus strategy

This section details the three steps of the SCS approach and provides the algorithms associated with each phase (Fig. 1).

(a) SCS-spectral, identifying graph sub-paths

The first phase of the SCS approach, called SCS-*spectral*, identifies subsets of vertices that are at a small *walk* distance from each other within the graph $\mathcal{G}$ (Fig. 1,a). This information is conveyed by the magnitude of the Laplacian eigenvector elements [51].

In the following, the input data matrix is $\mathbb{R}^{n\times p}$ with n the number of observations and p the number of variables. In Algorithm 1, the eigenvectors of the normalized Laplacian matrix L_{rw} are computed to identify vertices that lie on common sub-paths (Algorithm 1, lines $4-5$).

Algorithm 1: `SCS-spectral` → Path indicators

1 **In:** input data matrix $\mathbb{R}^{n\times p}$, with n the number of observations and p the number of variables
2 **Out:** U, normalized Laplacian eigenvector matrix
3 Compute $W \in \mathbb{R}^{p\times p}$, the mutual information matrix for the data points $\{x_i\}_{i\in[1,p]}$
4 Construct the unnormalized Laplacian $L = D - W$
5 Compute the **generalized** eigenvectors $\{v_k\}$ of the generalized eigenproblem $Lv = \lambda Dv$
6 Set U the matrix containing all the eigenvectors $\{v_k\}$ as columns in increasing order of $\{\lambda_k\}$.

In our consensus approach, we choose the mutual information to model vertex similarity as it provides a general measure of relationship between variables [56, 57]. Moreover, previous studies have shown that information theoretic measures are well suited to study high-dimensional biological data [58–60], which was one of our objectives when designing the `SCS` approach.

(b) SCS-learn, high-dimensional spectral embedding

The second phase of our approach, called `SCS`-*learn*, relies on the sign and magnitude of the first k eigenvector elements to reconstruct possibly overlapping sub-graphs that involve path connected vertices (Fig. 1,b). Specifically, each eigenvector v_k is associated with two sub-graphs, $\mathcal{G}_{v_k}^{m,-}$ and $\mathcal{G}_{v_k}^{m,+}$, that relate the m data points corresponding to either the most *negative* or the most *positive* eigenvector elements (Algorithm 2, line 7).

For clustering purposes, the subspace spanned by the first k eigenvectors would normally be preferred to their individual interpretation [28]. However the `SCS`-*learn* step does not aim at partitioning the variables, but rather to learn the whole underlying network based on overlapping sub-graphs. In particular, the non high-dimensional settings ($m \ll n$) obtained for each local reconstruction $\mathcal{G}_{v_k}^{m,+/-}$ restrict the number of *false positive* edges. Alternatively, the overlaps between selected subsets of m variables limit the number of *false negative* interactions. At the end of this phase, the edges eventually retained in each individual network $\mathcal{G}_l$ are those that were learned every time a sub-graph $\mathcal{G}_{v_k}^{m,+/-}$ involved the corresponding pair of vertices (Algorithm 2, lines $17-18$). Lastly, whenever the input reconstruction method $\mathcal{R}_l$ provides orientations, a *majority rule* is applied to set the final orientation or

Algorithm 2: `SCS-learn` → Embedded networks

1 **In:** U, first k eigenvector matrix (`SCS`-*spectral* output)
$\{\mathcal{R}_l\}$, a set of L network reconstruction methods
3 **Out:** $\{\mathcal{G}_l = (\mathbf{V}, \mathbf{E}_l)\}$, a set of L (possibly oriented) networks
4 **forall the** $\mathcal{R}_l$ *// in parallel* **do**
5 **forall the** v_k *with lowest* λ_k *// in parallel* **do**
6 Sort elements of $v_k = (v_k^1, ..., v_k^p)$ in increasing order
7 Using the $\mathcal{R}_l$ method, reconstruct two networks, $\mathcal{G}_{v_i}^{m,+}$ and $\mathcal{G}_{v_i}^{m,-}$, from the m most *positive* and m most *negative* elements of v_k
8 **for** *pairs* $(x_i, x_j) \in \mathbf{V}^2_{\mathcal{G}_{v_i}^{m,+/-}}$ **do**
9 $occurence_{\mathcal{G}_l}^{x_i x_j}$++
10 **if** (x_i, x_j) *adjacent in* $\mathcal{G}_{v_i}^{m,+/-}$ **then**
11 $adjacency_{\mathcal{G}_l}^{x_i x_j}$++
12 **if** $(x_i \leftarrow x_j$ *or* $x_i \rightarrow x_j)$ **then**
13 $orient_{\mathcal{G}_l}^{x_i x_j} \leftarrow orient_{\mathcal{G}_l}^{x_i x_j} \cup +/-1$
14
15
16 **end**
17 **end**
18 **end**
19 Set $\mathcal{G}_l = (\mathbf{V}, \mathbf{E}_l)$ for each $\{\mathcal{R}_l\}$ where $\mathbf{V} = \{x_i\}_{i\in[1,p]}$ and $\mathbf{E}_l = \{(x_i, x_j) | occurence_{\mathcal{G}_l}^{x_i x_j} = adjacency_{\mathcal{G}_l}^{x_i x_j}\}$
20 with $orient_{\mathcal{G}_l}^{x_i x_j} \leftarrow majority\{orient_{\mathcal{G}_l}^{x_i x_j}\}$

resolve possible conflicts over all the inferred orientations for two adjacent vertices (Algorithm 2, line 19). If no majority can be achieved, the edge is set undirected.

(c) SCS-consensus, final network

In this last phase, called `SCS`-*consensus*, networks inferred by individually embedded reconstruction methods are combined (Fig. 1,c). Specifically, for each learning approach $\mathcal{R}_l$, we rank the predicted edges by decreasing strength or confidence (Algorithm 3, lines $4-6$). Then, following the integration procedure proposed in [5], an average is computed to provide a consensus rank for the (x_i, x_j) edge in the final graph $\mathcal{G}$ (Algorithm 3, line 8). If an individual reconstruction method gives no edge between (x_i, x_j), the pair receives the worst possible rank for this method, i.e. $rank_{\mathcal{G}_l}^{x_i x_j} = 1$. A weighted average over the (sub)set $\{\mathcal{R}\}_{L'}$ of learning approaches that predicted orientations is also computed, giving greater weight to upper rank edge orientations (Algorithm 3, lines $9-12$). Lastly,

Algorithm 3: `SCS-consensus` → Final network

1 **In:** $\{\mathcal{G}_l = (\mathbf{V}, \mathbf{E}_l)\}$ for each $\{\mathcal{R}_l\}$ (`SCS-`*learn* output)
2 e_{max}, most significant edges threshold
3 **Out:** $\{\mathcal{G} = (\mathbf{V}, \mathbf{E})\}$, consensus (oriented) network
4 **forall the** $\mathcal{R}_l$ *// in parallel* **do**
5 Order $\{(x_i, x_j)\} \in \mathbf{E}_l$ by decreasing strength
6 **end**
7 **forall the** $(x_i, x_j) \in \mathbf{V}^2$ *// in parallel* **do**
8 $rank_{\mathcal{G}}^{x_i x_j} = \frac{1}{L}\sum_l^L \left(rank_{\mathcal{G}_l}^{x_i x_j}/|\mathbf{E}_l| \right)$
9 $orient_{\mathcal{G}}^{x_i x_j} = \frac{1}{\sum_l^{L'} w_{\mathcal{G}_l}^{rank(x_i x_j)}} \sum_l^{L'} orient_{\mathcal{G}_l}^{x_i x_j} w_{\mathcal{G}_l}^{rank(x_i x_j)}$
10 with
11 $w_{\mathcal{G}_l}^{rank(x_i x_j)} = (1 - rank_{\mathcal{G}_l}^{x_i x_j}/|\mathbf{E}_l|)$
12 $orient_{\mathcal{G}_l}^{x_i x_j} = 1$ if $x_i \to x_j$, -1 if $x_i \leftarrow x_j$
13 else 0
14 **end**
15 Set $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ where
16 $\mathbf{V} = \{x_i\}_{i \in [1,p]}$ and $\mathbf{E} = \{(x_i, x_j) | rank_{\mathcal{G}}^{x_i, x_j} \leq e_{max}\}$

only the e_{max} most significant edges are retained in the consensus network (Algorithm 3, line 15).

Results

The `SCS` approach embeds multiple reconstruction methods in a spectral framework to learn possibly oriented interactions from high-dimensional data by (i) combining the edges discovered from overlapping sub-graphs (Fig. 1, `SCS-`*learn*, (b)) and (ii) computing a consensus network (Fig. 1, `SCS-`*consensus*, (c)). In the following, the reconstructed networks are evaluated for an increasing proportion of eigenvectors (Fig. 2, horizontal axis). Results are discussed in terms of *Precision* ($TP/(FP+TP)$), *Recall* ($TP/(TP+FN)$) and *F-score* ($2\times Prec\times Rec/(Prec+Rec)$) ($FN$,$TP$,$FP$; *false negative*, *true positive* and *false positive* edges resp.). In particular, falsely oriented *TP* edges are considered as *FP*. For these evaluations, a benchmark network of 223 nodes and 338 edges has been considered (ANDES benchmark [61, 62]). This choice was in particular motivated by the fact that each variable of the ANDES benchmark network has exactly two categories, as encountered for metagenomics co-presence or presence-absence data. Besides, the 223 variables of this network enable us to reproduce high-dimensional conditions while evaluating the `SCS` results against reconstruction performed by each learning approach without the `SCS` embedding. We also considered a larger benchmark network composed of 1,041 nodes and 1,397 edges, MUNIN [63], and provide the corresponding results in (Additional file 1: Figures S3 and S9). We randomly sampled 5 datasets of sizes 150 and 200 to perform the experiments under high-dimensional conditions for `ANDES`, and 5 datasets of size 935 for `MUNIN`. The embedded reconstruction methods are `ARACNE` [1], a mutual information-based approach, `3off2` [4], a hybrid method that combines constraint-based and scoring approaches based on multivariate information measures, and a hill-climbing algorithm using the Bayesian Dirichlet equivalent score. We also considered a random classifier in our `SCS-`*spectral* and `SCS-`*learn* step evaluations (Additional file 1: Figures S4).

SCS-learn network evaluations

As previously established [5], adding high quality reconstruction methods to a consensus approach significantly improves consensus predictions. We have thus evaluated the accuracy improvement achieved in the `SCS-`*learn* phase that relies on the `SCS-`*spectral* step. Specifically, we have compared reconstructions obtained from variable subsets selected with the element magnitude of the first k eigenvectors to networks learnt based on variable subsets derived from different partitioning or clustering methods. Alternative subset selections are provided by spectral fuzzy C-means partitioning, spectral K-means clustering and recursive bi-partitioning. Random subset selection is also considered as a mere comparison.

Evaluations of embedded network reconstructions from subgraphs of $m = 12$ nodes using $n = 150$ samples (results for different subgraph and dataset sizes follow a similar trend, see Additional file 1) for the ANDES benchmark are given in Fig. 2 (top three rows). Reconstructions obtained from randomly sampled subsets exhibit a poor *Precision* (green solid line). This highlights that guided local reconstructions improve prediction accuracy. Networks reconstructed from subgraphs that rely on spectral K-means (darkblue solid line) or spectral fuzzy C-means (lightblue solid line) subsets do not provide better *Precision* than the `SCS-`*learn* method (red solid line) up to 30 eigenvectors (14 % of the total number). Although bipartition of the variables (salmon solid line) allows for better *Precision* than the random or spectral clustering, it is still largely outperformed by the `SCS-`*learn* phase.

This high *Precision* is at the slight expense of the *Recall* (Fig. 2, middle column), although it still outperforms the bi-partitioning approach and performs almost as better as clustering-based reconstructions. It is worth noting that reconstructions obtained with the `SCS-`*learn* step are consistent with Proposition 2 and 3. In particular, Fig. 2 shows an increase of the *Recall* as the number of eigenvectors grows (middle column, red solid line) as well as a higher *Precision* with the first eigenvectors (left column, red solid line). This is in line with a progressive discovery of the true underlying network and further show

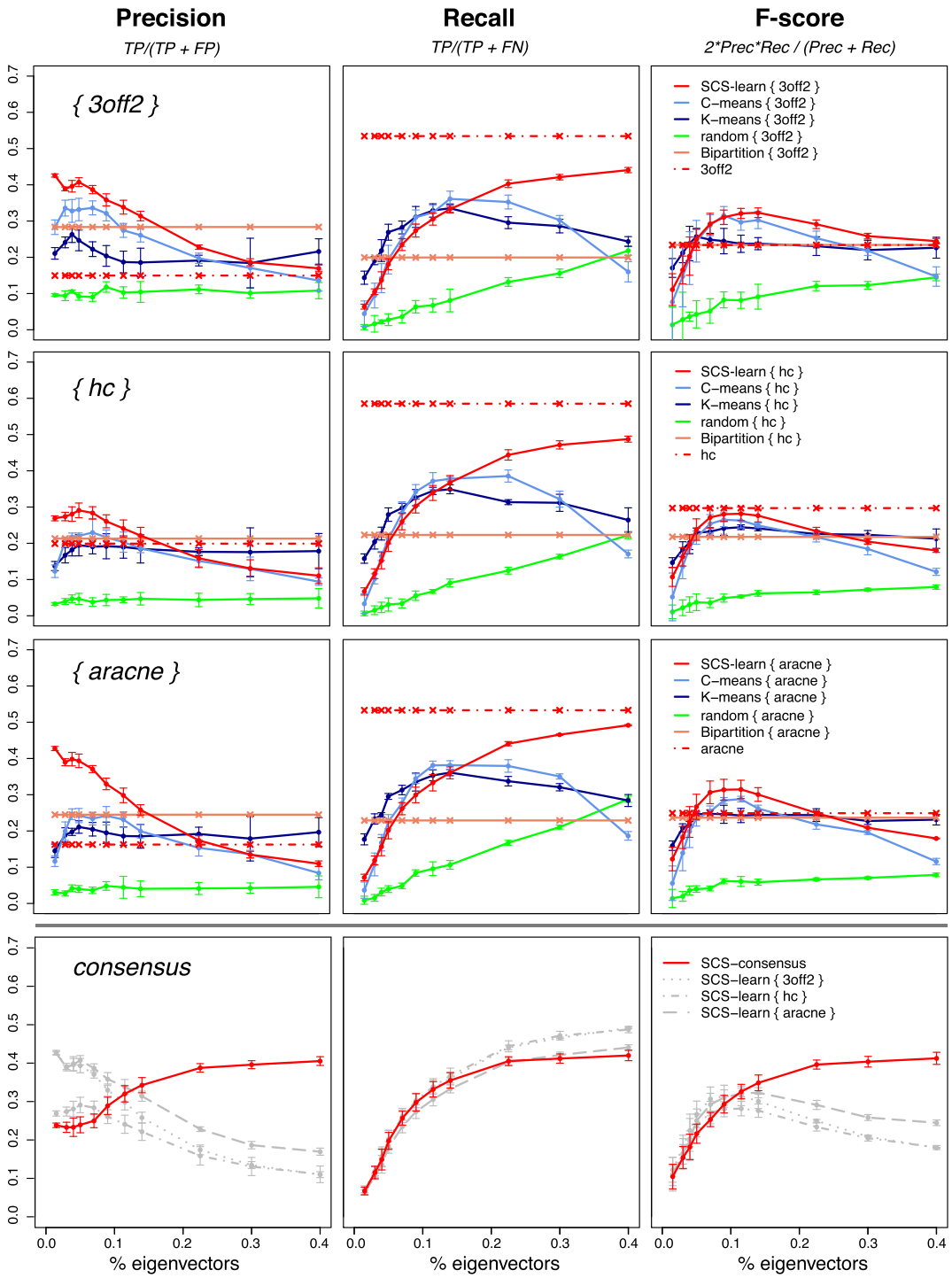

Fig. 2 SCS-learn and SCS-consensus evaluations for ANDES benchmark network [223 nodes, 338 edges, $\langle k \rangle = 3.03$]. *Precision*, *Recall* and *F-score* results for an increasing proportion of eigenvectors (up to 40 %), subgraphs of 12 nodes (5 % variables) and 150 samples. Scores take misorientations into account. Each point is an average over 5 datasets (results for different subgraph and dataset sizes follow a similar trend, see Additional file 1). (SCS-learn, top three rows) Three learning algorithms are embedded to reconstruct a network from subgraphs whose vertices are selected from the magnitude of eigenvector elements (SCS-*learn*, red solid line), spectral fuzzy C-means partitioning (*light blue* solid line), spectral K-means clustering (*dark blue* solid line), random subsets (*green* solid line) and recursive bi-partitioning (salmon solid line). Results are compared to scores obtained without spectral or partitioning embedding (*red* dashed line). (SCS-consensus, bottom row) The SCS-*learn* reconstructions are combined in a consensus network (*red* solid line) and compared with individual SCS-*learn* outcomes (*gray* dashed lines). Scores are computed from the top 338 consensus edges (results for different number of consensus edges follow a similar trend, see Additional file 1)

that non principal eigenvectors, although less informative than the first eigenvectors, carry relevant information on connected vertices. This can also be observed, to a lesser extent, when a random classifier is embedded in the SCS-*learn* step (Additional file 1: Figure S4). Interestingly, the *Recall* of networks based on spectral clustering partitions decreases when too many eigenvectors are considered (Fig. 2, middle column, lightblue and darkblue solid lines). As already established [26–29], truncated eigenvector basis are expected to emphasize variable similarities and thus, should indicate relevant variable subsets. Yet, due to the approximation error from the real valued solution, non principal eigenvectors are unreliable and worsen variable partitioning. Consequently, connected vertices may be assigned to distinct clusters as the number of eigenvector grows, leading to local reconstructions with a low *Recall*.

All together, the association of the SCS-*spectral* and SCS-*learn* steps leads to higher *F-score* results (Fig. 2, right column; Additional file 1: Figure S3, left column) as compared to reconstructions obtained with various partitioning approaches. This improvement is achieved from a relatively small number of eigenvectors (5 % of the total number), thus enabling a good trade-off between reconstruction quality and the number of required subgraphs. Lastly, the ANDES benchmark network was considered as its size allows for a direct reconstruction by each learning method. Results provided in Fig. 2 (dashed red line) show that SCS-*learn* performs better than, or as well as, reconstruction methods alone.

SCS-consensus network evaluations

Evaluations of consensus networks reconstructed from embedded learning approaches based on subgraphs of $m = 12$ nodes and using $n = 150$ samples are given in Fig. 2 (bottom row). The ANDES benchmark network having 338 edges, scores for the consensus outcome are given based on the 338 first ranked edges (results for different number of edges follow a similar trend, see Additional file 1). The consensus *Precision* scores (Fig. 2, bottom left, red solid line) clearly outperform the individually embedded learning approaches (gray dashed lines) as the proportion of eigenvector grows. Similar results are observed for the MUNIN benchmark network (Additional file 1: Figure S9).

Interestingly, these results emphasize the complementarity of the different reconstruction methods, as already demonstrated [5]. In particular, it has been shown that ARACNE and other mutual information reconstruction methods detect more easily feedfoward loop ($A \rightarrow B \rightarrow C$ and $A \rightarrow C$) and fan-in ($A \rightarrow C$ and $B \rightarrow C$) patterns. Conversely, cascade ($A \rightarrow B \rightarrow C$ and (A,B) not adjacent) and fan-out ($A \rightarrow C$ and $A \rightarrow B$) patterns are more easily inferred by Bayesian learning approaches [5].

All together, the SCS-*consensus* phase provides high *F-score* network reconstructions (Fig. 2, bottom right, red solid line) for a reasonable number of eigenvectors (proportion $\geq$ 11.5 %). The SCS-*consensus* predictions also exhibit high *F-scores* when considering variable subsets of larger sizes in the SCS-*learn* phase (Additional file 1: Figures S7–S9).

Reconstruction of microbial ecosystems

We applied the SCS method to a complex biological dataset generated by high-throughput sequencing of gut microbiome samples from 663 patients recruited in the MetaHIT project (Metagenomics of the Human Intestinal Tract). The nearly 4 million genes whose abundance was measured using quantitative metagenomics were binned to generate representative variables based on their mean co-abundance as introduced by Nielsen et al. [47]. These co-abundance groups (CAG) can be either classified as *genomic units* (GU) for small groups (between 3 and 700 genes) or *metagenomic species* (MGS) for larger groups (more than 700 genes). The authors produced a first reconstruction of the gut microbial ecosystem based on Fisher's exact test between pairs of CAGs.

In our study we used this extensively annotated dataset where information on phylogenetic classification and gene assembly is also available. Here we focused on $p = 2{,}101$ CAGs with more than 50 genes as already proposed in [64]. Figure 3a represents 307 co-presence relationships (edges) between these 2,101 CAGs (vertices) with at least one connection (leading to a subset of 445) as already provided by Nielsen et al. [47]. The number of genes composing a CAG is proportional to the vertex size. CAGs from the same phylum have similar color hues that are specified at the family level of their phylogenetic classification (e.g. *Firmicutes* are given in a range of blue and *Bacteroides* in a range of pink).

The SCS approach which embeds three reconstruction methods (ARACNE, 3off2, hill-climbing) inferred a consensus network of 6,389 edges from the above-mentioned dataset. To compare our results with the pairwise network reconstructed by Nielsen et al. [47], we selected the same number (307) of top-ranked SCS edges which represent approximately 5 % of the consensus interactions. This network composed of 443 nodes yields more complex substructural patterns as illustrated in Fig. 3b. When comparing networks (A) and (B), only 111 out of the 307 edges (36 %) inferred by Fisher's exact test are also predicted by the SCS method. Interestingly, 105 of these common edges (95 %) have genetic elements that share same assembly contigs, bringing strong biological evidence for these predicted relationships. Conversely, out of the remaining 196 edges solely inferred by Fisher's exact test, a significantly smaller number (121, 62 %) have genetic elements that share same assembly contigs ($p < 8 \times 10^{-10}$, χ^2).

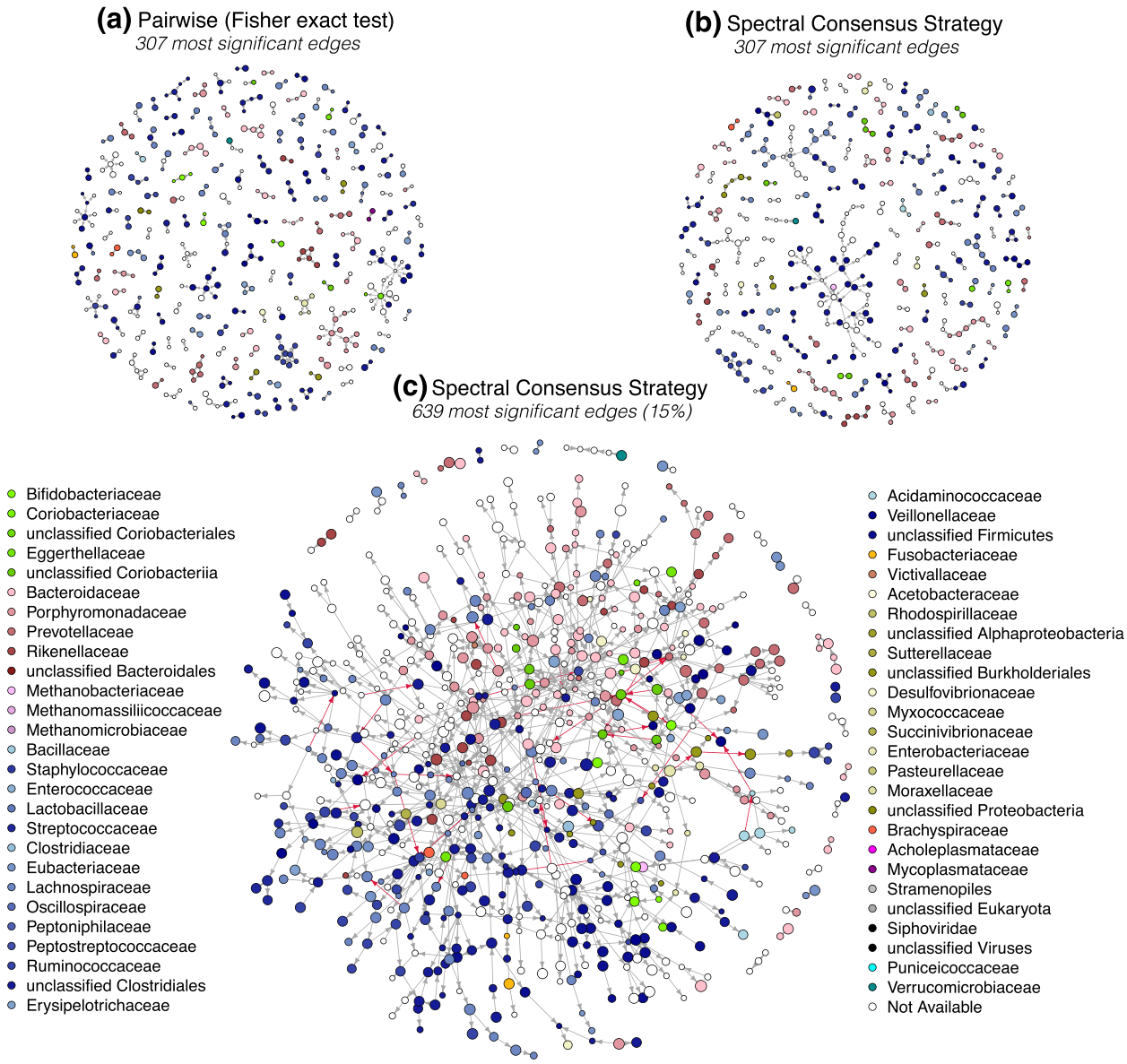

Fig. 3 Microbial co-presence ecosystem. Microbial ecosystem reconstructed with the pairwise Fisher's exact test [47] (**a**) and the SCS approach (**b,c**). Data for 2,101 co-abundant groups (CAGs) and $n = 663$ patients recruited in the MetaHIT project were used. Edges depict co-presence (*gray edges*) or absence-presence (*red edges*) relationships. **a** Gut microbial ecosystem based on Fisher's exact test between pairs of CAGs [47] (307 edges between 445 CAGs of at least 50 genes). **b** The same number of top-ranked edges (307) obtained with the SCS approach which involve 443 CAGs of at least 50 genes. **c** The 15 % most significant edges obtained with the SCS approach (654 nodes and 639 edges)

Complementary evaluations for different number of common edges (from 55 to 146 edges) follow the same trend (Additional file 1: Table S4 and Additional file 1: Figure S9). We hypothesize that a non negligible number of edges inferred by pairwise reconstruction techniques may correspond to indirect relationships.

We explored the topology of the SCS consensus gut microbial ecosystem at different most significant edges threshold (e_{max}) and illustrate the network at 15 % in Fig. 3c (654 vertices and 639 edges). The modular structure of this network is highlighted by tightly related vertices sharing similar colors. This indicates that species of the same family or phylum are mostly co-present as previously discussed [43]. This can be explained by the fact that closely related species have similar genetic background adapted for the same environmental niche. Of interest is also the fact that small CAGs (GU) are strongly linked with large CAGs (MGS) having the same phylogenetic annotations as depicted in Fig. 3(a & b) and previously described [47]. The SCS microbial network also

includes consensus directed edges computed from the orientations of the embedded 3off2 and hill-climbing algorithms. Gray oriented edges ($A \rightarrow B$) indicate *ordered* co-presence relationships (i.e. the presence of A species is expected whenever B is found). Conversely, red oriented edges provide presence-absence information.

We further analysed the SCS microbial network by considering the edge rank correlations between individual reconstructions and the consensus result (Fig. 4). The 3off2 and the ARACNE algorithms have a strong correlation (Fig. 4, $\rho = 0.77$), as it could be expected for approaches that rely on similar metrics. Conversely, the edge ranks between 3off2 or ARACNE and hill-climbing heuristic exhibit weak correlation coefficients (Fig. 4, $\rho = 0.31$ and $\rho = 0.22$ resp.). The slightly higher correlation between 3off2 and hill-climbing approaches may be related to the fact that 3off2 is a hybrid approach that is also score-based. All together, these results demonstrate the complementarity of the individual approaches from which the human gut microbial consensus predictions can benefit.

Discussion

In this paper, we propose a consensus network learning approach called *Spectral Consensus Strategy* which is based on spectral decomposition. Our method proceeds in three steps, namely SCS-*spectral*, SCS-*learn* and SCS-*consensus*. The first and second phases enable any reconstruction method to learn a possibly oriented network under high-dimensional settings. In addition to accuracy improvement of each reconstruction method, the spectral framework on which the SCS approach relies, also supports fast processing of high-dimensional datasets. The last phase combines the outcome of each reconstruction method to provide consensus predictions.

This strategy, as well as being accurate, scales up extremely well. Specifically, as the SCS-*learn* step processes in parallel local reconstructions related to the first k eigenvectors (Algorithm 2, lines $5 - 15$), it is the time complexity of the reconstruction methods that mainly impedes the whole running time. The SCS framework itself does not add any demanding computations. In particular, the running time for each individual reconstruction method embedded in the SCS-*learn* phase grows with the number of variables p as $\mathcal{O}(p \log p)$ (Algorithm 2, line 6). Furthermore, all reconstruction methods can simultaneously learn the network within the second phase. As an example, gut microbiota consensus reconstruction (2,101 variables, 663 samples, Fig. 3c) required 43 seconds to reconstruct all subgraphs ($m = 40$ vertices, 63 eigenvectors) needed for the $\mathcal{G}_l$ individual networks, and 52 seconds to build the consensus outcome $\mathcal{G}$ using 40 CPUs. Besides, the early step of the SCS-*spectral* phase which involves the computation of the mutual information matrix (Algorithm 1, line 3) and the last step of the SCS-*learn* phase which is dedicated to the assembling of local reconstructions (Algorithm 2, lines $17 - 19$), can be efficiently optimised and implemented [65, 66]. All together, the SCS approach could efficiently reconstruct the microbiome ecosystem, while the hill-climbing algorithm alone did not converge in 48 hours (see Additional file 1: Section 4, for detailed evaluations). These results highlight the ability of our method to improve the scalability of the embedded learning approaches.

The subgraph size m for the SCS-*learn* phase influences the quality of individual reconstructions ($\mathcal{G}_l$ graphs). Specifically, too small subgraphs lead to low *Recall* and very high *Precision*, while conversely too large subgraphs (even still under non high-dimensional conditions) increase the *Recall* at the expense of the *Precision*, both cases impeding the *F-score* results (Additional file 1: Figures S1–S3). Yet, predictions output by the SCS-*learn* step remain better than predictions derived from classical clustering and partitioning approaches for various sizes m. Interestingly, although the parameter m significantly impacts individual reconstructions, it only slightly impedes the consensus *F-score*. In particular, larger subgraphs still provide a consensus network of good quality from high-dimensional dataset (Additional file 1: Figures S7–S9). Similarly, the eigenvector proportion influences individual reconstructions $\mathcal{G}_l$ as too many eigenvectors lead to lower *Precision* and higher *Recall*. Yet, the consensus network based on the first e_{max} most significant edges achieves good and stable quality as the number of eigenvectors grows (Additional file 1: Figures S7–S9).

To define the minimal number of eigenvectors that would bring sufficient amount of information for a good consensus reconstruction, we designed a heuristic approach based on the decreasing interval between successive eigenvalues. For classical clustering approaches, the *eigengap* heuristic has been proposed to define the most suitable cluster number. This eigengap heuristic method is related to the fact that under ideal conditions, k distinct connected components are associated to the first k null eigenvalues and thus, a gap can be found between $\lambda_{i \leq k} = 0$ and $\lambda_{k+1} > 0$. In practice, the eigengap heuristic sets the number k such that $\lambda_{i \leq k}$ are small but λ_{k+1} is relatively large. The SCS approach objective is not to partition variables but rather to reconstruct a consensus network from overlapping subgraphs, using as much as possible of the information conveyed by each eigenvector. As shown from the counts and cumulative counts of *true positive* interactions for the ANDES benchmark network (Additional file 1: Figure S5), although most of the *true positive* interactions are retrieved from the first eigenvectors, non principal eigenvectors also conveyed relevant information

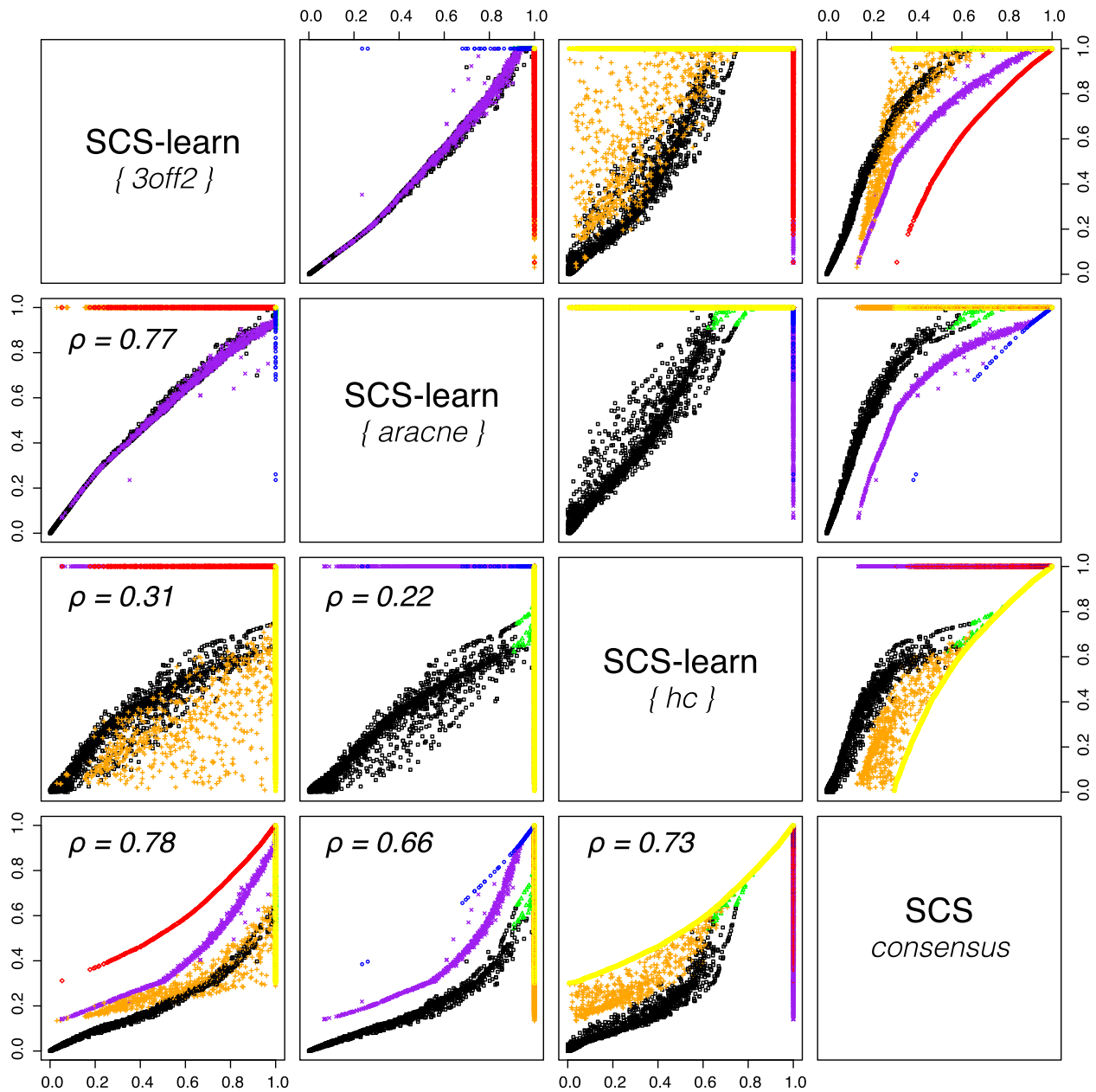

Fig. 4 Edge rank correlations between SCS-learn and SCS-consensus outcomes for human gut microbial ecosystem. 6,389 edges were predicted from a dataset of 663 observations and 2,101 CAGs (MetaHIT project [47]). Rank of edges predicted by only one embedded learning method are given in blue (`ARACNE`, 159 edges), red (`3off2`, 498 edges) and yellow (hill-climbing, 2,889 edges). Rank of edges predicted by two individual learning methods are given in green (`ARACNE` & hill-climbing, 31), orange (`3off2` & hill-climbing, 573 edges) and purple (`3off2` & `ARACNE`, 720 edges). Rank of edges predicted by all individual methods are given in black (1,519 edges)

on connected vertices. Hence, we consider the first k eigenvectors for which the successive eigenvalues are dissimilar enough as being the best number of eigenvectors to be used for the `SCS` consensus reconstruction. As an example, our heuristic method evaluated at 30 (14 %) the most suitable number of eigenvectors for the ANDES benchmark network. This number approximately corresponds to the number of eigenvectors from which the consensus network achieves better *F-score* results than networks obtained from individually embedded methods (Fig. 2).

The `SCS` approach is mainly designed to reconstruct large unknown biological networks, thus no weights have been assigned to individual reconstruction methods. However, if any prior knowledge is available on the underlying network topology, such as bias in particular

connection patterns, weights can be easily assigned when computing the average interaction rank.

Conclusion

Our contribution addresses the problem of large network reconstructions. The *Spectral Consensus Strategy* aims to reconstruct networks from high-dimensional dataset by overlapping subgraph parallel learning and consensus predictions. Although this approach is not intended to partition the data points, it takes advantage of spectral decomposition to identify tightly related vertices. We show by our experiments on both standard benchmark and real complex data that the performance of the proposed approach is extremely competitive. Our method is efficient from a computational viewpoint, its implementation is straightforward, and no effort has to be spent on hyper-parameter tuning.

Additional file

Additional file 1: Contains complementary demonstrations as well as supplementary evaluations for the SCS approach. Specifically, Section 1 provides Propositions and associated *sketches of proof* that support our method. Complementary evaluations of the SCS first steps, namely SCS-*spectral* and SCS-*learn*, are given in Section 2. We also provide in Section 3 complementary evaluations of the SCS last step, named SCS-*consensus*. Execution time comparisons are given in Section 4. Supplementary statistics on the application to human gut microbiota close this Additional file 1. (PDF 606 kb)

Declarations

This article has been published as part of *BMC Bioinformatics* Volume 17 Supplement 16, 2016: Proceedings of the Tenth International Workshop on Machine Learning in Systems Biology (MLSB 2016). The full contents of the supplement are available online at http://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-16.

Funding

The publication of this article was funded by a grant from the AP-HP to JDZ (Contrat d'interface APH-HP, 2014). This work is also supported by the European Union's Seventh Framework Program under grant agreement HEALTH-F4-2012-305312 (MetaCardis project). SA was financed by an ICAN-Danone Research grant.

Availability of data and materials

The MetaHIT data we use in our experiments have been already published as a supplementary material of Nielsen H.B., al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature biotechnology 32(8), 822–828 (2014). Available from: doi:10.1038/nbt.2939, and the supplementary material can be provided.

Authors' contributions

SA, NS, EP and JDZ conceived and performed the research. SA, NS, EP and JDZ wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Integromics, Institute of Cardiometabolism and Nutrition, ICAN, Assistance Publique Hôpitaux de Paris, Pitié-Salpêtrière Hospital, Paris 75013, France. [2]Sorbonne Universités, UPMC University Paris 6, UMR S U1166 NutriOmics Team, Paris 75013, France. [3]UMR S U1166 Nutriomics Team, INSERM, Paris 75013, France. [4]IRD, UMI 209, UMMISCO, IRD France Nord, Bondy F-93143, France.

Published: 13 December 2016

References

1. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006:7(Suppl 1).
2. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):8.
3. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. In: International Conference on Computational Molecular Biology. New York: Mary Ann Liebert, Inc.; 2000. p. 601–20.
4. Affeldt S, Verny L, Isambert H. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. BMC Bioinformatics. 2016;17(S-2):12.
5. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9(8):796–804.
6. Smet RD, Marchal K. Advantages and limitations of current network inference methods. Nat Rev Micro. 2010;10(8):717–29.
7. Bellot P, Olsen C, Salembier P, Oliveras A, Meyer PE. Netbenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference. BMC Bioinformatics. 2015;16(312):1–15.
8. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci. 2010;107(14):6286–291.
9. Pearl J, Verma TS. A theory of inferred causation. Studies in Logic and the Foundations of Mathematics. 1995;134:789–811.
10. Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Soc Sci Comput Rev. 1991;9:62–72.
11. Cooper GF, Herskovits E. A bayesian method for the induction of probabilistic networks from data. Mach Learn. 1992;9(4):309–47.
12. Heckerman D, Geiger D, Chickering DM. Learning bayesian networks: the combination of knowledge and statistical data. Mach Learn. 1995;20(3):197–243.
13. Cano A, Gomez-Olmedo M, Moral S. A score based ranking of the edges for the PC algorithm. In: Proceedings of the Fourth European Workshop on Probabilistic Graphical Models; 2008. p. 41–8.
14. de Campos L. A scoring function for learning bayesian networks based on mutual information and conditional independence tests. J Mach Learn Res. 2006;7:2149–187.
15. Tsamardinos I, Brown L, Aliferis CF. The max-min hill-climbing bayesian network structure learning algorithm. Mach Learn. 2006;65(1):31–78.
16. Fiedler M. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czechoslov Math J. 1975;25(100):619–33.
17. Fiedler M. Algebraic connectivity of graphs. Czechoslov Math J. 1973;23(98):298–305.
18. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell. 2000;22(8):888–905.
19. Spielman DA, Teng SH. Spectral partitioning works: Planar graphs and finite element meshes. In: Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on. IEEE; 1996. p. 96–105.
20. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci. 2006;103(23):8577–82.
21. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psych. 1933;24:417–41.
22. Kruskal JB, Wish M. Multidimensional scaling. Beverely Hills: Sage Publications; 1978.

23. Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 1998;10(5):1299–319.
24. Azar Y, Fiat A, Karlin A, McSherry F, Saia J. Spectral analysis of data. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing. ACM; 2001. p. 619–626.
25. Kannan R, Vempala S, Vetta A. On clusterings: Good, bad and spectral. J ACM. 2004;51(3):497–515.
26. Perona P, Freeman WT. A factorization approach to grouping. In: European Conference on Computer Vision. Springer; 1998. p. 655–70.
27. Alpert C, Kahng A, Yao S. Spectral partitioning: the more eigenvectors, the better. Discrete Appl Math. 1999;90:3–26.
28. Ng AY, Jordan MI, Weiss Y. On Spectral Clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems. MIT Press; 2001. p. 849–56.
29. Brand M, Huang K. A Unifying Theorem for Spectral Embedding and Clustering. In: Proc. 9th International Workshop on AI and Statistics; 2003. http://www.merl.com/publications/TR2002-042/.
30. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol. 2008;26(10):1155–60.
31. Fröehlich H, Klau GW. Reconstructing consensus bayesian network structures with application to learning molecular interaction networks. In: GCB. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik; 2013. p. 46-55. http://dx.doi.org/10.4230/OASIcs.GCB.2013.46.
32. Berto S, Perdomo-Sabogal A, Gerighausen D, Qin J, Nowick K. A consensus network of gene regulatory factors in the human frontal lobe. Front Genet. 2016;7:.
33. Lancichinetti A, Fortunato S. Consensus clustering in complex networks. Scientific reports. 2012;2:.
34. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65.
35. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, Pettersson S. Host-gut microbiota metabolic interactions. Science. 2012;336(6086):1262–7.
36. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489(7415):220–30.
37. Walsh CJ, Guinane CM, O'Toole PW, Cotter PD. Beneficial modulation of the gut microbiota. FEBS Lett. 2014;588(22):4120–30.
38. Clarke SF, Murphy EF, Nilaweera K, Ross PR, Shanahan F, O'Toole PW, Cotter PD. The gut microbiota and its relationship to diet and obesity: new insights. Gut Microbes. 2012;3(3):186–202.
39. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: human gut microbes associated with obesity. Nature. 2006;444(7122):1022–3.
40. Elson CO, Cong Y. Host-microbiota interactions in inflammatory bowel disease. Gut Microbes. 2012;3(4):332–44.
41. Lepage P, Häsler R, Spehlmann ME, Rehman A, Zvirbliene A, Begun A, Ott S, Kupcinskas L, Doré J, Raedler A, et al. Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. Gastroenterology. 2011;141(1):227–36.
42. Bajaj JS, Thacker LR, Heuman DM, Fuchs M, Sterling RK, Sanyal AJ, Puri P, Siddiqui MS, Stravitz RT, Bouneva I, et al. The stroop smartphone application is a short and valid method to screen for minimal hepatic encephalopathy. Hepatology. 2013;58(3):1122–32.
43. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014;513(7516):59–64.
44. Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, Stonebraker AC, Hu C, Wong FS, Szot GL, Bluestone JA, et al. Innate immunity and intestinal microbiota in the development of type 1 diabetes. Nature. 2008;455(7216):1109–13.
45. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed F. Gut metagenome in european women with normal, impaired and diabetic glucose control. Nature. 2013;498(7452):99–103.
46. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55–60.
47. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32(8):822–8.
48. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–50.
49. Mohar B, Alavi Y, Chartrand G, Oellermann O. The laplacian spectrum of graphs. Graph Theory Comb Appl. 1991;2(871-898):12.
50. Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007;17(4): 395–416.
51. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E. 2006;74(3):.
52. Golub GH, van Loan CF. Matrix Computations. Johns Hopkins Series in the Mathematical Sciences. Favoritenstrasse: The Johns Hopkins University Press; 1989.
53. Pothen A, Simon HD, Liu K.-P. P. Partitioning sparse matrices with eigenvectors of graphs. Technical report NASA Ames Research Center. 1989.
54. Kleinberg JM. Authoritative sources in a hyperlinked environment. J ACM (JACM). 1999;46(5):604–32.
55. Miller B, Bliss N, Wolfe PJ. Subgraph detection using eigenvector L1 norms. In: NIPS; 2010. p. 1633–1641. http://www.bibsonomy.org/bibtex/22fa92e5556307d62c4ed6473f4bba10c/dblp.
56. Russakoff DB, Tomasi C, Rohlfing T, Jr CRM. Image similarity using mutual information of regions. In: 8th European Conference on Computer Vision (ECCV. Springer; 2004. p. 596–607.
57. Liu R, Gillies DF. An eigenvalue-problem formulation for non-parametric mutual information maximisation for linear dimensionality reduction. In: International Conference on Image Processing, Computer Vision, and Pattern Recognition; 2012. p. 905–910.
58. Priness I, Maimon O, Ben-Gal IE. Evaluation of gene-expression clustering via mutual information distance measure. BMC Bioinformatics. 2007;8:.
59. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics. 2002;18:231–40.
60. Butte AJ, Kohane IS, Kohane IS. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput. 2000;5:415–26.
61. Scutari M, Denis JB. Bayesian Networks with Examples in R. Boca Raton: Chapman and Hall; 2014.
62. Conati C, Gertner AS, VanLehn K, Druzdzel MJ. On-line student modeling for coached problem solving using Bayesian networks. In: User Modeling. Springer; 1997. p. 231–42.
63. Andreassen S, Jensen F, Andersen S, Falck B, Kjærulff U, Woldbye M, Sørensen A, Rosenfalck A, Jensen F. MUNIN - an Expert EMG Assistant. In: Computer-Aided Electromyography and Expert Systems, Chapter 21. Noth-Holland: Elsevier; 1989.
64. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500(7464):541–6.
65. Sales G, Romualdi C. parmigene-a parallel r package for mutual information estimation and gene network reconstruction. Bioinformatics. 2011;27(13):1876–7.
66. Qiu P, Gentles AJ, Plevritis SK. Fast calculation of pairwise mutual information for gene regulatory network reconstruction. Comput Methods Prog Biomed. 2009;94(2):177–80.