Ensemble-based network aggregation improves the accuracy of gene network reconstruction

Jeffrey D. Allen[1,2], Yang Xie[1,2] and Guanghua Xiao[1,*]

[1] Quantitative Biomedical Research Center
[2] Simmons Cancer Center, UT Southwestern Medical Center,

* To whom correspondence should be addressed. Tel: 214-648-4553; Fax: 214-648-7673; Email: Guanghua.Xiao@UTSouthwestern.edu

Present Address: Guanghua Xiao, 5323 Harry Hines Blvd, UT Southwestern Medical Center, Dallas, TX, 75390, USA

ABSTRACT

Reverse engineering approaches to construct context-specific gene regulatory networks (GRNs) based on genome-wide mRNA expression data have led to significant biological findings. However, the reliability and reproducibility of the reconstructed GRNs needs to be improved. Here, we propose an ensemble-based network aggregation approach to improve the accuracy of the network topology constructed from mRNA expression data. To evaluate the performance of different approaches, we created dozens of simulated networks and also tested our methods on three Escherichia coli datasets. We demonstrate three novel applications from this development. First, bootstrapping can be done on the available samples, turning any network reconstruction approach into an ensemble method. Second, this aggregation approach can be used to combine GRNs from different network inference methods, creating a novel network reconstruction approach that consistently outperforms any constituent method. Third, the approach can be used to effectively integrate GRNs constructed from different studies – producing more accurate networks. We are releasing an implementation of these techniques as an R package "ENA" which is able to run network inference in parallel across multiple servers. We made all of the code and data used in our simulations and analysis available online at https://github.com/QBRC/ENA-Research to ensure the reproducibility of our results.

INTRODUCTION

Accurate reconstruction of Gene Regulatory Networks (GRNs) from gene expression microarrays has been shown to be valuable in a myriad of areas surrounding biomedical research[1–5]. Researchers have previously used approaches including Bayesian Network-Based approaches [6], Correlation-Based approaches [7], and Partial-Correlation-Based approaches [8,9]. These methods have been shown to have various strengths and weaknesses under different biological/simulation settings, with no one method excelling in all conditions. [10]. Additionally, leveraging gene expression data from multiple datasets to construct gene networks is often difficult, due to discrepancies in microarray platform selection, as well as in normalization and data processing techniques. In this study, we propose an Ensemble-based Network Aggregation (ENA) approach to integrate gene networks derived from different methods and different datasets to improve the accuracy of network inference.

We used a non-parametric, inverse-rank-product, algorithm in the ENA approach to combine networks reconstructed on the same set of genes. The rank- product method was introduced by

PeerJ PrePrints | https://peerj.com/preprints/40v1/ | v1 received: 3 Jul 2013, published: 3 Jul 2013, doi: 10.7287/peerj.preprints.40v1

Breitling et al [11,12] as an effective method for detecting differentially expressed genes in microarray studies. Because the rank product method is powerful and computationally efficient, it was extended to be used in other fields, such as RNAi screening [13] and proteomics [14]. This method can be directly related to the linear rank statistics [15]. In this study, we show three ways to leverage this approach to generate the ensemble-based networks: 1.) Samples in a dataset can be "bootstrapped" to reconstruct multiple networks out of a single original dataset using a single reconstruction method, which can then be aggregated into a more accurate and reproducible network; 2.) Networks produced by various reconstruction methods can be aggregated into a single network that is more accurate than the network provided by any individual method; 3.) Networks reconstructed from different studies which contain the same genes can be combined into a single, more accurate network, despite differences in platforms or normalization techniques. Because this approach has little overhead, it can efficiently be applied to dozens or hundreds of networks reconstructed on the same set of genes. We find that this approach has the ability to improve the accuracy of GRN reconstruction in all three applications based on simulated gene expression data, as well as *Escherichia coli* (E. coli) datasets[16–19].

MATERIAL AND METHODS

Overview of the Inverse-Rank-Product Network Aggregation Approach

Reconstructed gene networks are often returned as a weighted undirected graph $G = (N, \Omega)$, where G is a reconstructed graph, $N = \{1, ..., n\}$ is the set of vertices (genes) in the graph, and $\Omega = [\omega_{ij}]_{i,\, j \in N}$ is referred to as the adjacency matrix, in which ω_{ij} represents the intensity of the interaction between genes i and j. A larger (absolute) value of ω_{ij} indicates a stronger interaction or higher confidence in the edge between genes i and j, while $\omega_{ij} = 0$ indicates no interaction, or conditional independence between genes i and j. Some techniques, such as Sparse PArtial Correlation Estimations (SPACE) [9], return a sparse matrix in which many of the possible interactions are 0; other techniques return complete graphs in which all edges are present with non-zero weightings. Additionally, the distribution of ω_i can vary drastically among reconstruction techniques. For this reason, the aggregation of networks reconstructed using different techniques or different datasets is challenging. In this study, we used a rank-product method to combine networks to overcome the different distributions observed in this problem.

Specifically, suppose $G = \{G^k\}$ is a set of networks constructed on the same set of genes N, where $k = \{1, ..., K\}$ is the index of a particular network. For each single network $G^k = (N, \Omega^k)$, we calculate r_{ij}^k, the rank of ω_{ij}^k for $\{i, j \in N \text{ and } i < j\}$. Since the adjacency matrix Ω of an undirected graph is a symmetric matrix, we only need to calculate the rank of the $N*(N-1)/2$ elements in ω_{ij} constituting the lower triangle (i < j) of the adjacency matrix. In this study, we give the lower rank to the high strength/confidence interaction. For example, the interaction with the highest strength/confidence will have rank 1. This operation is performed on each individual graph G^k independently. After the rank of r_{ij}^k has been computed for each network G^k , we calculate the rank of a particular edge between genes i and j in the aggregated network by taking the product of the ranks of the same edge across all networks in $\mathcal{G}$, as follows: $\tilde{r}_{ij} = \prod_{k=1}^{K} r_{ij}^k$ .This function is iterated over all possible edges to construct the aggregated network $\tilde{G} = (N, \tilde{r}_{ij})$, in which the strength of the edges in the new network are based on the aforementioned rank-product calculation.

This algorithm can be efficiently applied to large networks with many reconstructed networks in $\mathcal{G}$. The complexity of the algorithm is $O\left(K \cdot |N| \log\left(|N|\right)\right)$, as $\frac{|N|^2 - |N|}{2} = O(N^2)$ elements must be sorted for each network in G^k .

Three Applications for Ensemble-based Network Aggregation

The initial application was to leverage the rank-product method to “bootstrap” samples. Each time, we construct the gene network using a randomly selected subset of the available samples. By repeating this process B times, we create a set $\mathcal{G}$ consisting of B graphs, each reconstructed using only randomly selected bootstrap samples in the dataset. For example, here is the procedure to generate the bootstrapping network from the microarray dataset MD:

$$MD \xrightarrow{Bootstrap} \left\{ \begin{array}{lllll} MD^1 & \rightarrow & G^1 = \{N, \Omega^1\} & \rightarrow & r_{ij}^1 \ (\text{for } 1 \le i < j \le n) \\ \vdots & & \vdots & & \vdots \\ MD^B & \rightarrow & G^B = \{N, \Omega^B\} & \rightarrow & r_{ij}^B \ (\text{for } 1 \le i < j \le n) \end{array} \right\} \rightarrow \text{RankProduct} \rightarrow \tilde{G}$$

Of course, this bootstrapping procedure inflates the computational complexity of GRN reconstruction by orders of magnitude, as GRNs must be reconstructed B times, rather than just once. Because each graph in $\mathcal{G}$ can be reconstructed independently, it is possible to take advantage of the “parallelizability” of these simulations by utilizing multiple cores or computers as we discuss below. Note also that the complexity of GRN reconstruction does scale on the order of samples included, so each permuted GRN can be constructed slightly more quickly than a single global GRN; for the reconstruction techniques employed in this study, however, the performance did not vary greatly based on the number of samples included.

The second application of the rank-product network merging method was to reconstruct an aggregated GRN based on the output of multiple different reconstruction techniques. We have observed that reconstruction techniques perform differently based on different simulation settings [20], with no one method outperforming the others on all metrics. Thus, we were interested to see whether or not merging these GRNs would offer an improvement in performance. In this application, the set of graphs G consist of one graph per network reconstruction technique employed. In our analysis, we leveraged GeneNet[8], Weighted Correlation Network Analysis (WGCNA) [7], and Sparse PArtial Correlation Estimation (SPACE) [9], creating a set of 3 graphs which can then be aggregated. GeneNet and SPACE are partial-correlation-based inference algorithms. GeneNet uses the Moore-Penrose pseudoinverse [21] and bootstrapping to estimate the concentration matrix. The SPACE algorithm creates a regression problem when trying to estimate the concentration matrix and then optimizes the results with a symmetric constraint and an L1 penalization. WGCNA is a correlation-based approach which can identify sub-networks using hierarchical clustering. Conceptually, the aggregated graph would place higher confidence on those edges which were consistently ranked highly across the three methods, and would place lower confidence on those edges which were only ranked highly in one graph. This is the procedure to derive the ensemble network based on M different methods on the same dataset MD:

$$MD\left\{\begin{array}{llll} \xrightarrow{\text{method 1}} & G^1=\{N,\Omega^1\} & \to & r_{ij}^1\ (\text{for } 1\le i<j\le n) \\ \vdots & \vdots & & \vdots \\ \xrightarrow{\text{method M}} & G^M=\{N,\Omega^M\} & \to & r_{ij}^M\ (\text{for } 1\le i<j\le n) \end{array}\right\} \to \text{RankProduct} \to \tilde{\text{G}}$$

The final application evaluated in this study was in the merging of networks constructed from different datasets. Historically, gene expression datasets have been collected from various sites on different microarray platforms with different procedures for tissue collection; this creates incompatibilities and difficulties when trying to perform analysis on data from different datasets simultaneously. Because the rank-product method makes no assumptions on the distribution of the data at any point, we employ it to combine GRNs produced from different datasets, yielding a single, aggregated GRN which aims to capture the consistencies in network topology from the GRNs produced on different datasets. Here is the procedure to derive the aggregated network from datasets MD^1, MD^2…. MD^D:

$$\left.\begin{array}{lllll} MD^1 & \to & G^1=\{N,\Omega^1\} & \to & r_{ij}^1\ (\text{for } 1\le i<j\le n) \\ & \vdots & \vdots & & \vdots \\ MD^D & \to & G^D=\{N,\Omega^D\} & \to & r_{ij}^D\ (\text{for } 1\le i<j\le n) \end{array}\right\} \to \text{RankProduct} \to \tilde{\text{G}}$$

Software

PeerJ PrePrints | https://peerj.com/preprints/40v1/ | v1 received: 3 Jul 2013, published: 3 Jul 2013, doi: 10.7287/peerj.preprints.40v1

The code used to bootstrap samples and aggregate the resultant networks was written in the R programming language [22]. We created an R Package entitled “ENA” and have made it available on CRAN (http://cran.r-project.org/web/packages/ENA/index.html); the compiled binaries, as well as all original source code are available for download there.

Because of the parallelization opportunities in this algorithm, we ensured that our software would be able to distribute the bootstrapping process across multiple cores and multiple nodes using MPI [23]. Thus, if 150 CPU cores were available simultaneously, a bootstrapping of 150 samples could run in approximately the same amount of wall-clock time as a single reconstruction using all samples could. The ENA package includes robust documentation and (optionally) leverages the RMPI package to allow parallel execution of the bootstrapping simulations where such a computational infrastructure is available.

Additionally, we leveraged the Git revision control system via GitHub (http://github.com) to control not only the R code developed for the ENA package, but also all code, reports, and data used in the aforementioned simulations and reconstruction techniques; all of this code is freely available at https://github.com/QBRC/ENA-Research. All the data analysis code that has been used to generate the results in this study was compiled into a single report and can be reproduced easily by using the knitr R package [24]. Due to the computational complexity involved in reconstructing this quantity of gene regulatory networks, the execution may take quite some time when analyzing the larger networks if not distributed across a large compute cluster.

RESULTS

Simulation

We first tested the ENA methods on a wide array of simulated datasets. We simulated the gene expression datasets based on previously observed protein-protein interaction networks[25,26], and the expression data were simulated from conditional normal distributions [27]. We extract five different network sizes in an approximately scale-free topology: 17 genes with 20 connections, 44 genes with 57 connections, 83 genes with 114 connections, 231 genes with 311 connections, or 612 genes with 911 connections. For each network size, we simulated datasets with differing numbers of samples (microarrays): 20, 50, 100, 200, 500, and 1,000. Finally, we varied the noise by setting the standard deviation of the expression values to either 0.25, 0.5, 1.0, or 1.5. In total, we generated 120 datasets to cover all possible arrangements of the above variables.

To test the effect of integrating networks derived from different datasets, we generated three different datasets, each containing 200 samples, from the 231-gene networks with noise values (standard deviation of the distribution of gene expression) of 0.25, 1, and 2. We then used the methods described above to reconstruct three networks, one from each dataset and then aggregate those networks. For comparison, we also combined all three datasets into a single dataset containing these 600 samples and reconstructed a single network from this larger dataset.

The performance of methods in this setting can be represented on a Receiver Operating Characteristic (ROC) Curve, which plots the True Positive Rate against the False Positive Rate, demonstrating the performance of the method at all relevant edge weight thresholds. The performance of a method can be quantified by calculating the Area Under the ROC Curve (AUC). The greater the AUC, the better the performance of the method represented. A perfect reconstruction would have an AUC of 1, and a random guess could obtain an AUC of 0.5.

Ensemble networks derived from bootstrapping Samples

We found that bootstrapping samples can increase the accuracy of network inference. For example, the networks reconstructed from the dataset on the 231-gene network with a noise value of 0.25 can be compared to demonstrate the variations in performance as seen in Figures 1 and 2. Figure 1 shows that by bootstrapping samples in the SPACE algorithm, the AUC of the reconstructed network can improve from 0.75 to 0.82. Figure 2 shows the degree of AUC improvement with each iteration of bootstrapping on SPACE, WGCNA and GeneNet with sample sizes of 20, 50 and 100 (left, middle and right panels). From this figure, the bootstrapping method increases the performance of SPACE substantially, improves GeneNet slightly when the number of microarrays is small, but does not noticeably improve the performance of WGCNA. SPACE benefits from bootstrapping in 80% of all simulated networks, and in 89% of “large” network simulations. Figure 3 shows the average performance increase achieved by bootstrapping SPACE on different network sizes. The improvement increases as the network size increases. Based on this evidence, we suggest employing the bootstrapping approach when using the SPACE algorithm, but not the others evaluated in this study.

Ensemble networks derived from different methods

Aside from optimizing individual reconstruction techniques, we find that combining different network reconstruction techniques executed on the same dataset also has the power to significantly improve the accuracy of the reconstructed networks. Using the dataset from the 83-gene network with 200 samples and a noise value of 0.25, we can review the comparative performance of each reconstruction technique, as well as the aggregated network. Figure 4 shows that the aggregated network outperforms any of the individual reconstruction techniques.

We observe that this trend holds true across most of the datasets that we tested: the aggregated method typically outperforms any single reconstruction technique. This is especially beneficial in scenarios in which the top performing individual network reconstruction technique may vary based on the context – some methods perform well on larger networks, others excel in datasets containing few samples, etc. To have an aggregation technique which consistently outperforms or matches the best performing individual method eliminates the need to choose a single reconstruction technique based on the context.

Ensemble networks derived from different datasets

Finally, we find that the ENA approach works very well when attempting to integrate various datasets, especially among heterogeneous datasets that contain different distributions of expression data. After generating three datasets from the 231-gene network, each with 200 samples and noise values of 0.25, 1, and 2, we reconstruct each network using Bootstrapped SPACE, GeneNet, and WGCNA, then aggregate the resultant networks into a single network for each dataset, producing one aggregated network for each of the three datasets. We then use the ENA approach to consolidate these three networks into a single network representing the underlying network behind the three distinct datasets. We compare this to the alternative of simply merging all three datasets into a single 600-sample dataset and using the same approach to reconstruct a single network. As shown in Figure 5, we find the proposed ENA approach outperforms the alternative approach of simply combining the expression data into a single dataset. Reconstructing on each dataset independently produces AUCs of 0.96, 0.96, and 0.89 for noise values of 0.25, 1, and 2, respectively. Naïvely merging the datasets by combining them into one large dataset yields an AUC of 0.96. The network aggregation approach yields the best performance, with an AUC of 0.98.

Evaluating ENA approach in E. coli datasets

We then tested the ENA approach on three Escherichia coli (E.coli) datasets: 1. The Many Microbe Microarrays Database (“M3D”)[16] contains 907 microarrays measured under 466 experimental conditions using Affymetrix GeneChip E.coli Genome arrays. 2. The second dataset (“Str”) is expression data from laboratory evolution of Escherichia coli on lactate or glycerol (GSE33147)[17]. This dataset contains 96 microarrays measured under laboratory adaptive evolution experiments using Affymetrix E. coli Antisense Genome Arrays. 3. The third dataset [18,19] (“BC”) contains 217 arrays measuring the transcriptional response of E.coli to different perturbations and stresses, such as drug treatments, UV treatments and heat shock. The RegulonDB database[28,29] contains the largest and best-known information on transcriptional regulation of E.coli and was used as the gold standard to evaluate the accuracy of constructed networks.

We were able to obtain similarly positive results by employing these approaches on the E coli data. Bootstrapping and aggregating the three methods on each dataset independently produced AUCs of 0.574, 0.616, and 0.599 for the BC, Str, and MD3 datasets respectively. By merging the three networks produced on each dataset using ENA, we were able to produce a network with an AUC of 0.655, larger than the AUC of any network produced by any of the datasets independently.

DISCUSSION

The ability to aggregate networks using the rank-product merging approach has shown to be a valuable contribution in reconstructing gene regulatory networks – and likely to other fields, as well. By bootstrapping a single dataset using a single approach such as SPACE, we were able to significantly improve the performance of the algorithm. By aggregating the networks produced by different reconstruction techniques on a single dataset, we are able to consistently match or outperform the best-performing technique for that dataset, regardless of fluctuations in the performance of any one algorithm. By aggregating networks constructed independently on different datasets capturing similar biological environments, we are able to reconstruct the network more accurately than would be possible using any one dataset alone.

It is likely that SPACE was the only method to show consistent and significant improvement from bootstrapping because the SPACE algorithm models the gene regulation using linear regression; as a result, the network construction problem is converted to a variable selection problem. In SPACE, the variable selection problem is solved by sparse regression techniques with a symmetric constraint. By solving all the regression models simultaneously, SPACE is trying to get the globally optimized results. However, due to the instability in variable selection [30] caused by collinearity in the data, the networks constructed by SPACE are sensitive to sampling. A small change in the samples selected may lead to a relatively large change in the network structure. As a result, the networks constructed from bootstrapping samples are relatively “independent”, which leads to better accuracy in the aggregated network.

We provide a user-friendly R package to allow others to use these techniques on their own datasets. By leveraging the MPI framework, we are able to run the bootstrapping process in parallel across many cores and nodes, drastically reducing the amount of time it takes to run such analysis. We include in this package a function which can permute random networks and perform ENA in order to better estimate the significance of any particular connection observed in a network. This can be used to reduce a continuous, complete graph to an unweighted graph including only statistically significant edges.

Finally, we went to great lengths to ensure that all of our analysis would be as reproducible as possible by structuring our analysis code in reproducible reports – most of which can be regenerated at the click of a button – and making all of these freely available online at https://github.com/QBRC/ENA-Research. We feel that this transparency is an important but uncommon step in the scientific process and hope that other researchers begin incorporating such practices in their investigation to foster more open, collaborative research.

Availability

The R code used to perform all of the analysis contained in this study is available in the R package entitled “ENA,” available on CRAN currently. The source code, as well as compiled binaries, are available for download at http://cran.r-project.org/web/packages/ENA/.

REFERENCES

1. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science’s STKE 303: 799–805.

2. Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science (New York, NY) 302: 249–255. doi:10.1126/science.1087447.

3. Segal E, Shapira M, Regev A, Pe D, Botstein D, et al. (n.d.) Module Networks : Discovering Regulatory Modules and their Condition Specific Regulators from Gene Expression Data. Cell Research.

4. Sachs K, Perez O, Pe’er D, Lauffenburger D a, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science (New York, NY) 308: 523–529. doi:10.1126/science.1105809.

5. Lee I, Date S V, Adai AT, Marcotte EM (2004) A Probabilistic Functional Network of Yeast Genes. Science 306: 1555–1558.

6. Friedman N, Linial M, Nachman I, Pe’er D (2000) Using Bayesian networks to analyze expression data. Journal of computational biology : a journal of computational molecular cell biology 7: 601–620. Available: http://www.ncbi.nlm.nih.gov/pubmed/11108481.

7. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics 9: 559. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2631488&tool=pmcentrez&rendertype=abstract. Accessed 12 July 2012.

8. Schäfer J, Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics (Oxford, England) 21: 754–764. Available: http://www.ncbi.nlm.nih.gov/pubmed/15479708. Accessed 12 July 2012.

9. Peng J, Wang P, Zhou N, Zhu J (2007) Partial Correlation Estimation by Joint Sparse Regression Model: 1–52.

10. Langfelder P, Luo R, Oldham MC, Horvath S (2011) Is my network module preserved and reproducible? PLoS computational biology 7: e1001057. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3024255&tool=pmcentrez&rendertype=abstract. Accessed 11 June 2011.

11. Breitling R, Herzyk P (2005) Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. Journal of bioinformatics and computational …. Available: http://www.worldscientific.com/doi/abs/10.1142/S0219720005001442. Accessed 22 February 2013.

12. Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS letters 573: 83–92. Available: http://www.ncbi.nlm.nih.gov/pubmed/15327980. Accessed 4 February 2013.

13. Birmingham A, Selfors L, Forster T (2009) Statistical methods for analysis of high-throughput RNA interference screens. Nature Methods 6: 569–575. Available: http://www.nature.com/nmeth/journal/v6/n8/abs/nmeth.1351.html. Accessed 22 February 2013.

14. Wiederhold E, Gandhi T, Permentier HP, Breitling R, Poolman B, et al. (2009) The yeast vacuolar membrane proteome. Molecular & cellular proteomics : MCP 8: 380–392. Available: http://www.ncbi.nlm.nih.gov/pubmed/19001347. Accessed 22 February 2013.

15. Koziol J (2010) Comments on the rank product method for analyzing replicated experiments. FEBS letters 584: 941–944. Available: http://www.sciencedirect.com/science/article/pii/S0014579310000542. Accessed 22 February 2013.

16. Faith JJ, Driscoll ME, Fusaro V a, Cosgrove EJ, Hayete B, et al. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic acids research 36: D866–70. doi:10.1093/nar/gkm815.

17. Fong SS, Joyce AR, Palsson BØ (2005) Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome research 15: 1365–1372. doi:10.1101/gr.3832305.

18. Sangurdekar DP, Srienc F, Khodursky AB (2006) A classification based framework for quantitative description of large-scale microarray data. Genome biology 7: R32. doi:10.1186/gb-2006-7-4-r32.

19. Xiao G, Wang X, Khodursky AB (2011) Modeling Three-Dimensional Chromosome Structures Using Gene Expression Data. Journal of the American Statistical Association 106: 61–72. doi:10.1198/jasa.2010.ap0950.Modeling.

20. Allen JD, Xie Y, Chen M, Girard L, Xiao G (2012) Comparing Statistical Methods for Constructing Large Scale Gene Networks. PLoS ONE 7: e29348. Available: http://dx.plos.org/10.1371/journal.pone.0029348. Accessed 19 January 2012.

21. Penrose R (1954) A Generalized Inverse for Matrices. Proc Cambridge Phil Soc 51: 406–413.

22. R Core Team (2012) R: A Language and Environment for Statistical Computing.

23. Gabriel E, Fagg GE, Bosilca G, Angskun T, Dongarra JJ, et al. (n.d.) Open MPI : Goals , Concept , and Design of a Next Generation MPI Implementation.

24. Xie Y (2012) knitr: A general-purpose package for dynamic report generation in R. Available: http://cran.r-project.org/package=knitr.

25. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, et al. (2004) Human protein reference database as a discovery resource for proteomics. Nucleic acids research 32: D497–501. doi:10.1093/nar/gkh070.

26. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, et al. (2006) Human protein reference database--2006 update. Nucleic acids research 34: D411–4. doi:10.1093/nar/gkj141.

27. Pan W, Lin J, Le CT (2002) Model-based cluster analysis of microarray gene-expression data. Genome biology 3: RESEARCH0009.

28. Salgado H, Martı´nez-Flores I, Lo´pez-Fuentes A, Garcı´a-Sotelo JS, Liliana Porro´ n-Sotelo HS, et al. (2012) Extracting Regulatory Networks of Escherichia coli from RegulonDB. In: Helden J, Toussaint A, Thieffry D, editors. Bacterial Molecular Networks. New York, NY: Springer New York, Vol. 804. pp. 179–195. doi:10.1007/978-1-61779-361-5.

29. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz-Rascado L, et al. (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic acids research 39: D98–105. doi:10.1093/nar/gkq1110.

30. Breiman L (1996) Heuristics of Instability and Stabilization in Model Selection. The Annals of Statistics 24: 2350–2383.

TABLE AND FIGURES LEGENDS

Figure 1. Receiver Operating Characteristic (ROC) curves demonstrating the performance of the SPACE algorithm on the 231-gene network with 20 samples and a noise value of 0.25 when performing a single iteration or bootstrapping the dataset using the Ensemble Network Aggregation approach. In this case, the Area Under the ROC Curve (AUC) of the non-bootstrapped SPACE method is 0.748, while the bootstrapped SPACE method is 0.816.

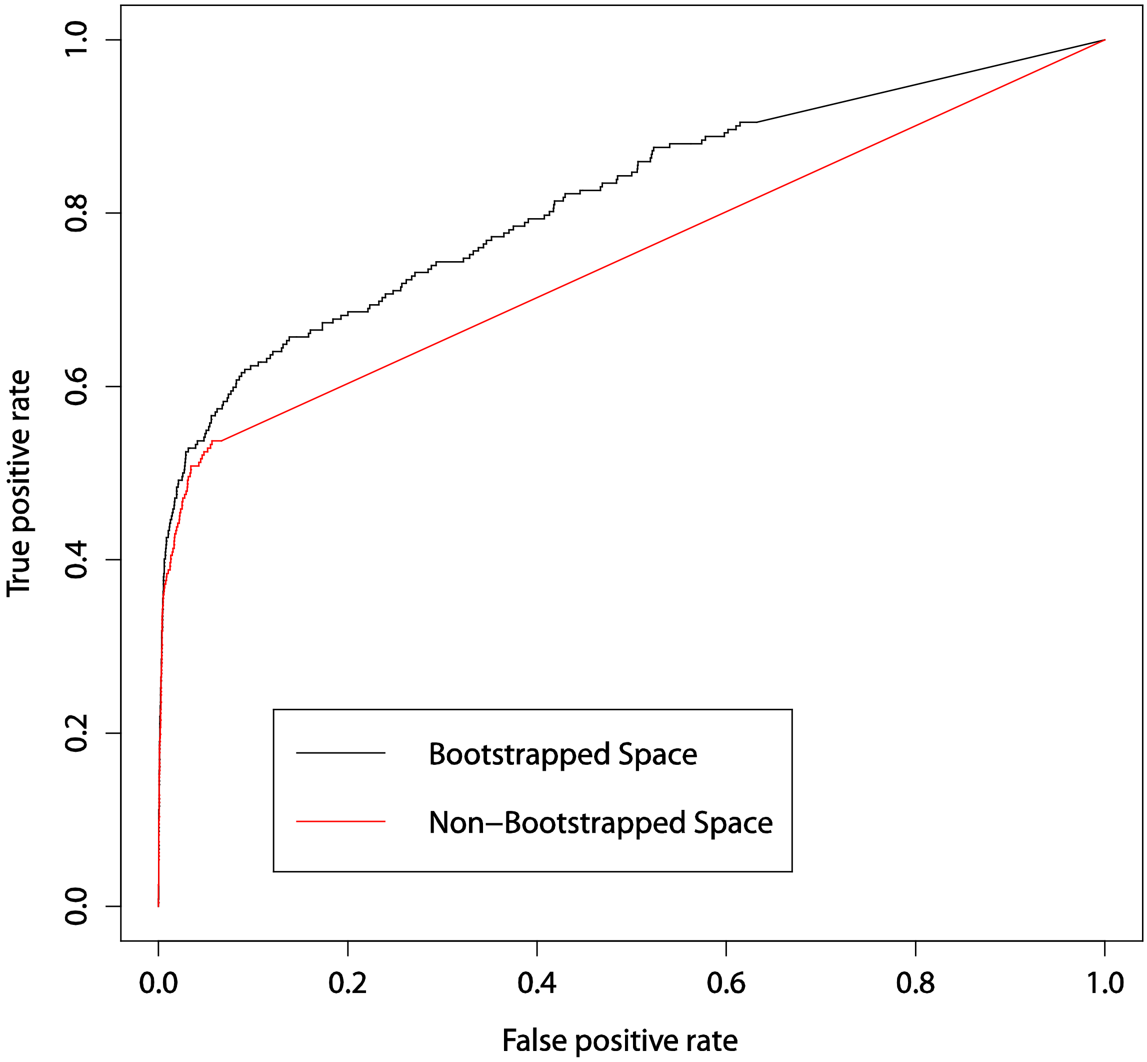


Figure 2. Comparison of the AUCs of the reconstructed networks from the 231-gene network with a noise value of 0.25 and different sample sizes (20, 50 or 100) for SPACE (a.), GeneNet (b.), and WGCNA(c.). In these plots, the y-axis shows the performance of the reconstructed network, measured by the Area Under the Curve; a horizontal line is drawn to represent the AUC of the non-bootstrapped reconstruction (a single reconstruction using all available samples). The x-axis represents the number of iterations in the bootstrapping process. Points below the horizontal line represent a loss in accuracy of the reconstructed networks, and points above the horizontal line represent a gain of AUC – an increase in performance.

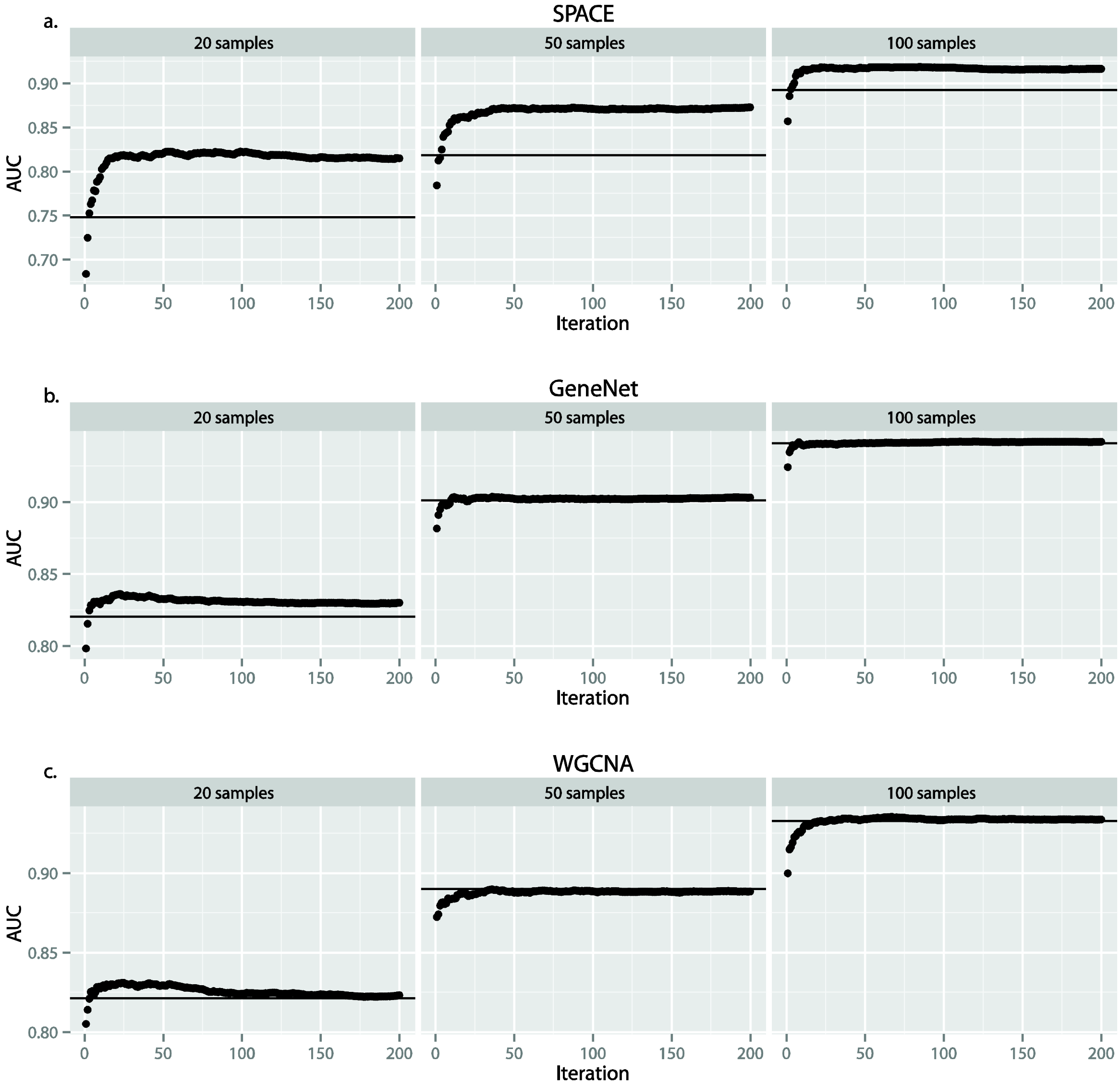


Figure 3. The effect of network size on ENA performance. The y-axis represents the improvement in AUC of the bootstrapped SPACE networks vs. the non-bootstrapped SPACE networks. Different bars represent different sizes of networks in the simulation study.

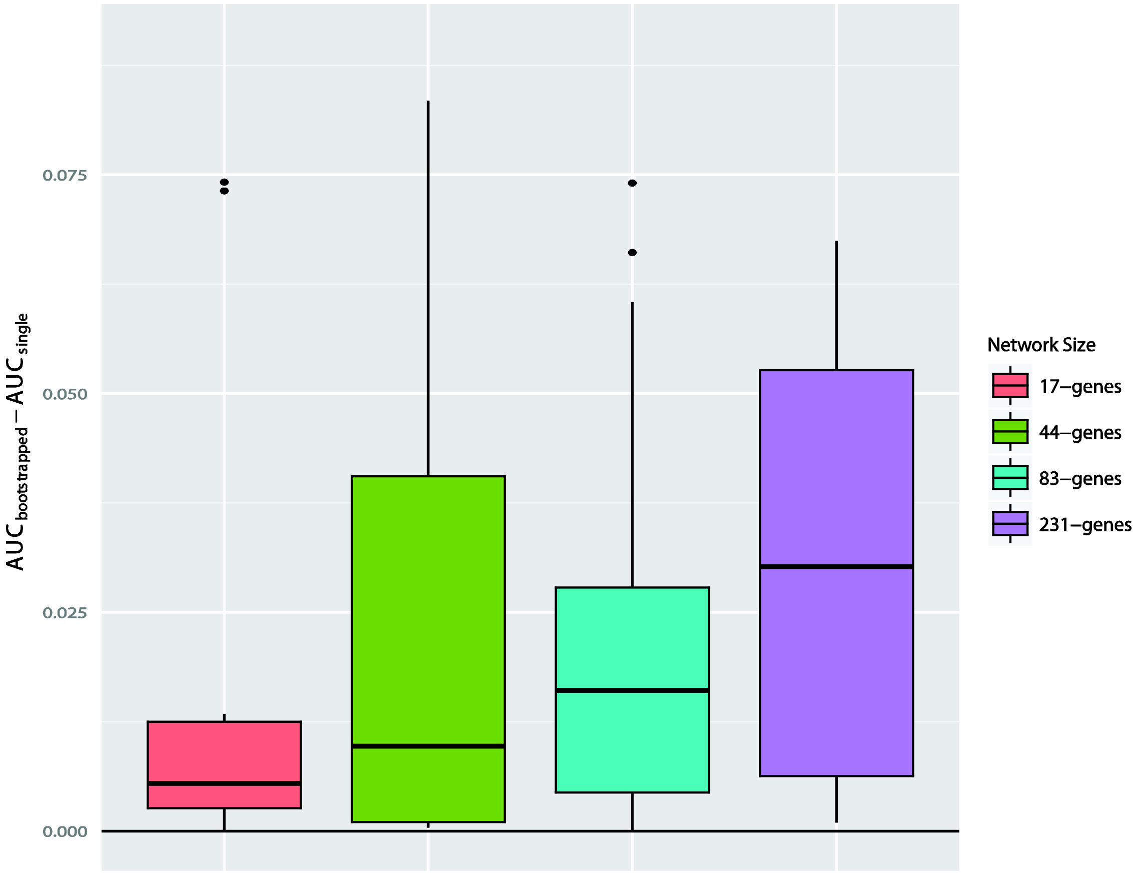


Figure 4. The performance of aggregating different methods. A comparison of the accuracy of the reconstructed networks using the dataset containing 200 samples (left) and 1,000 samples (right) from the 83-gene network with a noise value of 0.25. As can be seen, the ensemble network aggregation approach performs better than any of the other individual techniques on these two networks.

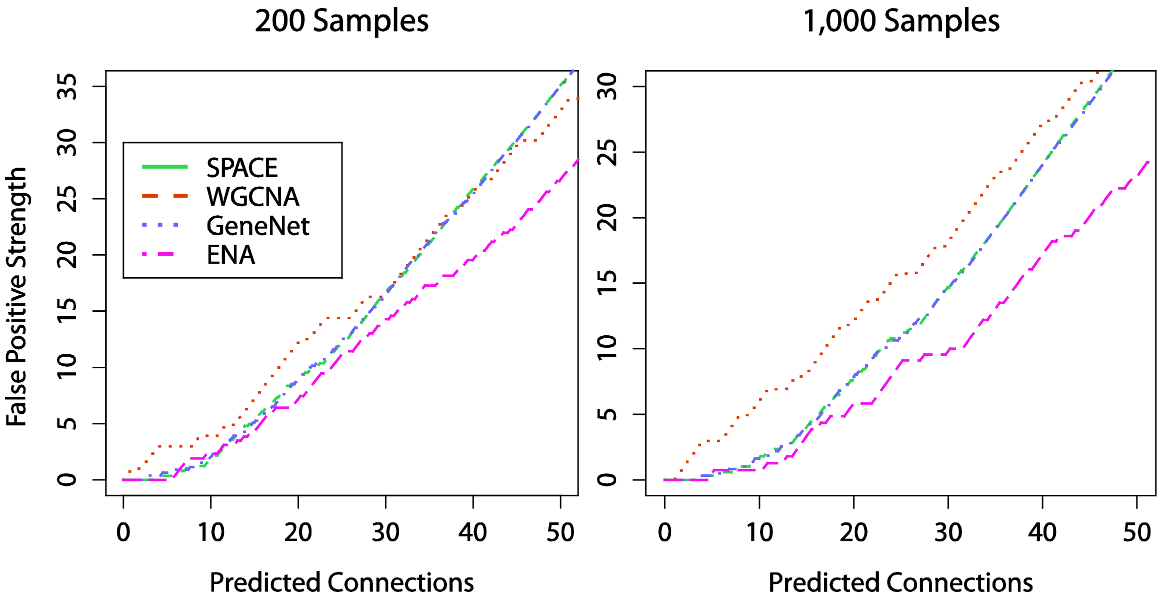


Figure 5. The ROC curves of different approaches to reconstruct the gene network based on three simulated datasets.

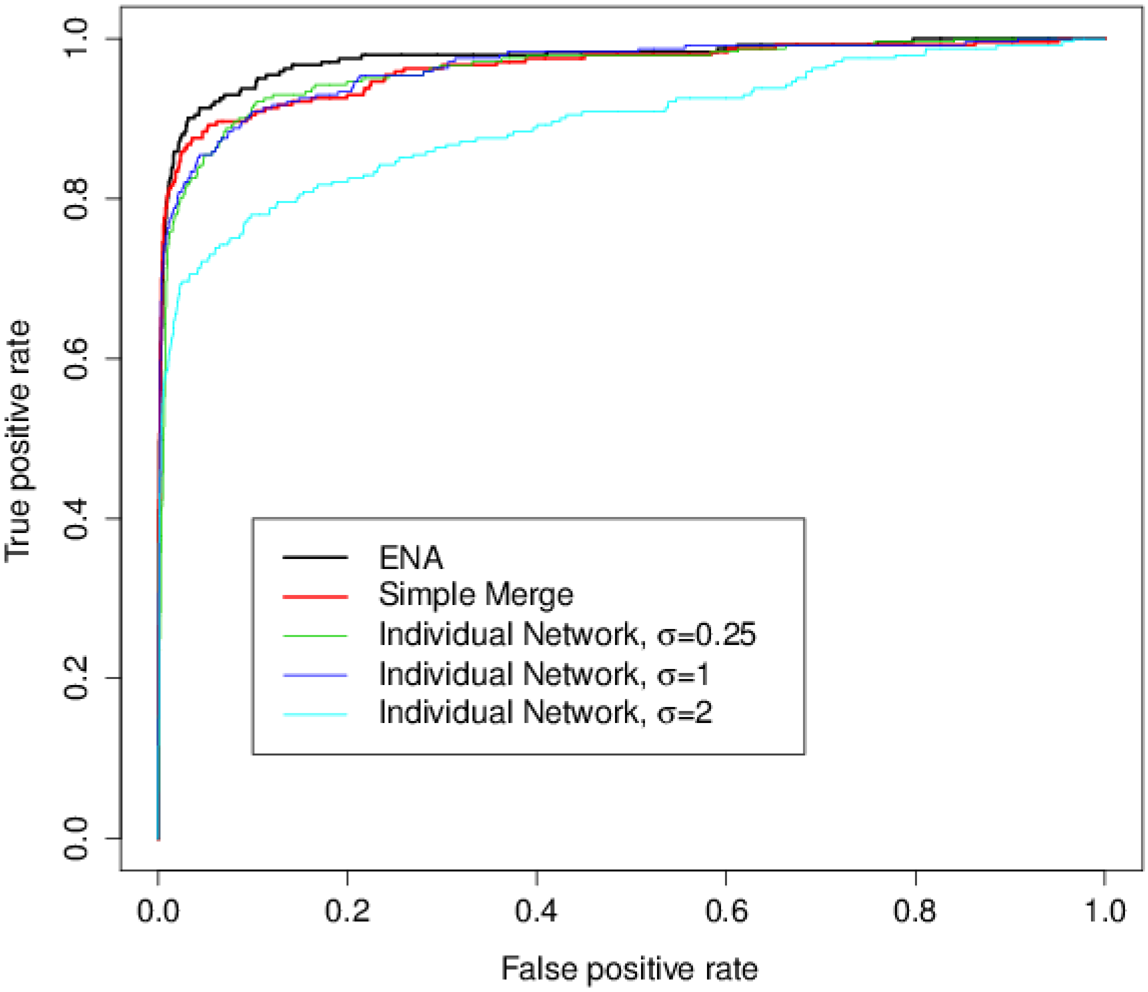
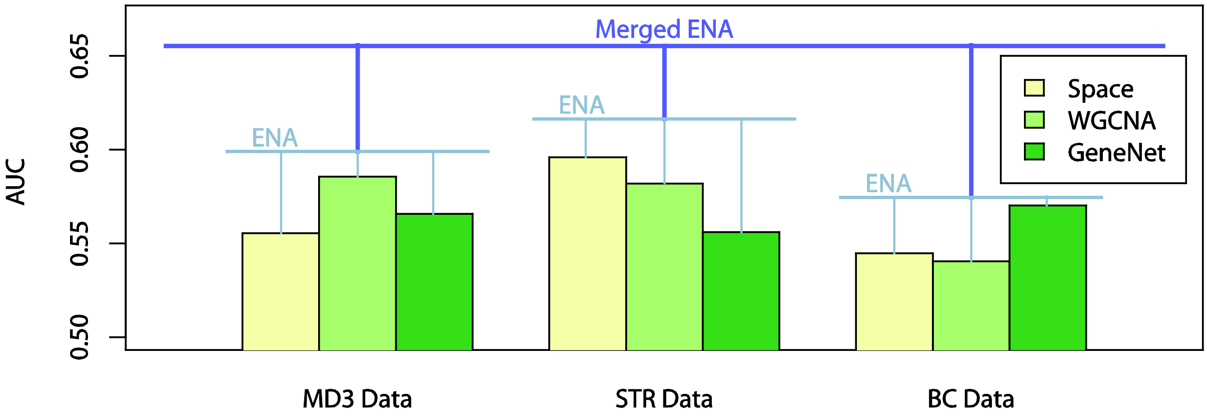


Figure 6. The AUCs of the produced networks when executing on the E. coli datasets. Note that the aggregating networks from SPACE, WGCNA and GeneNet increases the accuracy within each individual dataset, then aggregating results from three datasets further increases the accuracy beyond what any one dataset offered.



Supplementary File ENA-master.zip - contains the source code for the ENA R package,

Supplementary File ENA-Research-Master.zip - contains all of the reproducible analysis code behind this manuscript.