hSNIM: Hyper Scalable Network Inference Machine for Scale-Free Protein-Protein Interaction Networks Inference

Muhammad Shoaib Sehgal SHOAIB.SEHGAL@GMAIL.COM
Computer Science Department
Stanford University
Stanford, CA 94305

Abstract

The Protein-Protein Interaction (PPI) networks play an important role in cellular functionality. Due to their importance PPI networks have gained a wide spread attention from the research community. The PPI networks have various practical applications in biology in general, and in target drug discovery in particular. The PPI networks are, however, mostly unknown and inference of these biological networks using machine learning methods is far from trivial. In the past few years several machine learning models are developed to infer these networks. Despite the significant contributions, these models either have high *False Positive Rates* or they don't scale to infer networks from real world high dimensional biological data, most of which contains high number of missing values. In this report we'll present *Hyper Scalable Network Inference Machine* (hSNIM) to infer large scale PPI networks. The model not only scales to infer human interactome (human PPI network) but also can provide the best accuracy compared to the existing methods. The model achieves this accuracy by fusing various forms of data like Gene Expression, Partial PPI data, Sub-cellular localization and Phylogenetic trees. The experimental results corroborate that hSNIM has highest AUC 0.82 compared to existing proposed models like *Kernel Metric Learning* (KML), *Kernel Canonical Correlation Analysis* (KCCA) and *Matrix Completion with EM* (MCEM) when applied to infer Yeast interactome. The hSNIM model achieved 98% accuracy when it was used to infer human PPI network, which shows its ability to infer large scale networks accurately.

Keywords: Biological Networks, Protein-Protein Interaction Networks, Structure Learning and Large Scale Data Fusion.

1 Introduction

Proteins interact to form cellular functionality and the complete repertoire of these interaction is called *Protein-Protein Interaction* (PPI) network. From biological standpoint PPI networks not only help to elucidate the biological process and provide holistic picture of the biological processes, but also have huge applications in biology for instance, in tailored drug discovery.

A PPI network can be viewed as a graph $G\{V,E\}$ with set of vertices $V\{v_1,v_2, \ldots, v_n\}$ and edges $E\{e_1,e_2,\ldots e_m\}$ [1]. Each node in the graph represents a protein and the edge between a pair of proteins represents interaction. The objective is to fully infer G from the given experimental data (See Figure 1). To do so one of the approaches applied by biologists today is to do *in vivo* experiments to find out interacting proteins. Once the interaction is observed under certain experimental conditions, it is reported in the literature in the form of research publications. The process is very time consuming and costly since usually this type of experiments usually, at best, find out interacting proteins in the scale of tens. However, the total proteins in human are estimated to be in millions and we don't know the exact number today.

To address this problem several *Machine Learning* (ML) models have been proposed to infer PPI networks. The models range from simple classification models to very sophisticated versions of Bayesian Networks and Kernel Methods.

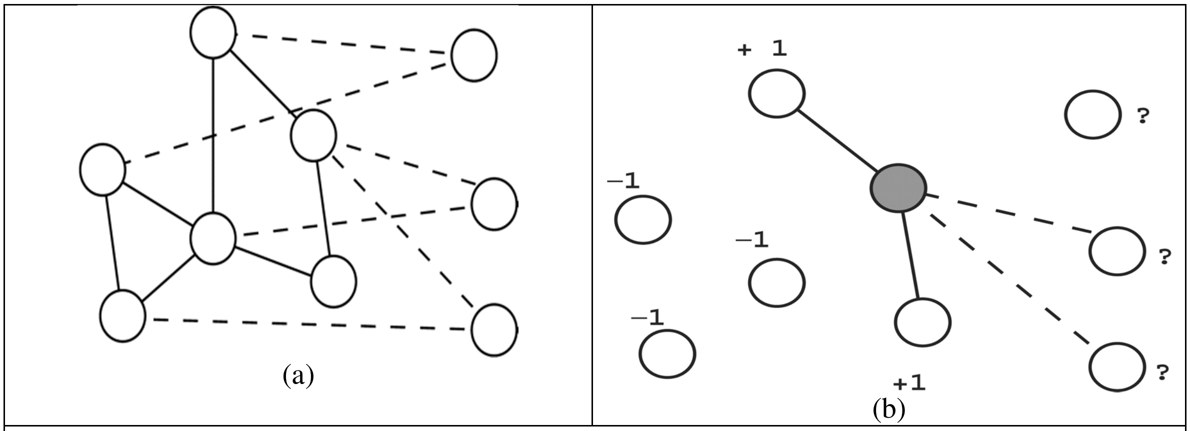

Figure 1. (a) Protein-Protein interaction network where vertices represent proteins and edges represent interaction. The dotted and solid lines denote inferred and manually curated edges respectively. Absence of edges between the vertices shows that there was no evidence of the interaction between the proteins. **(b)** The PPI network inference can be considered as a binary classification y = {+1,-1} problem where +1 means edge exists between vertices and -1 is for the vertices where edge doesn't exist. The objective is to find label for the protein pairs labelled ? in the figure (Image sources: K. Bleakley *et al.* [1]).

Despite all these efforts, however, the accuracy of the inferred PPI networks is still needs to be improved to gain the confidence of biologists. The low accuracy of the predicted PPI networks can be attributed to, but not limited to:

1- Inherent noise and variation in the biological data

2- Lack of available training data

3- High dimensionality

4- Missing values

5- Data has number of features *m* >> number of samples *n*

6- Complexity of the biological systems

Despite all these challenges, the accuracy of the predicted networks can be significantly improved if ML models could make use of all the information available. Generally PPI information could be inferred from the Gene expression data, *In vivo* PPI experimental data, *Sub Cellular Localization* (SCL) and phylogenetic profiles and in some cases from Y2H screens.

This report will address some of aforementioned problems by introducing a new ML model, *Hybrid Scalable Network Inference Machine* (hSNIM). The hSNIM achieves high accuracy compared to the existing models by accurately fusing the information present in the Gene expression, partial PPI network information, SCL, and phylogenic profiles data. The rationale behind our proposed approach is that each set of the data captures part of the picture and thus by merging the classification outcome from each data will help us to infer more accurate PPI networks. The model hSNIM model has shown not only to scale to infer human interactome unlike existing models present in the literature but also has demonstrated improved accuracy when compared against existing models like *Kernel Metric Learning* (KML), *Kernel Canonical Correlation Analysis* (KCCA) and *Matrix Completion with EM* (MCEM).

In addition as alluded to earlier, biological data contains missing data. The problem is more prevalent when different types of experimental data like SCL data, Gene expression data, and phylogenetic profiles are combined together. It is therefore, imperative to use a model which should be able to infer networks in the presence of missing information. Later in this report we'll show that hSNIM could capture human interactome in the presence of missing information unlike aforementioned competitive models.

Rest of the report consists of following Sections. Section 2 presents an overview of existing techniques while detail architecture and theory behind hSNIM is presented in the Section 3. Section 4 provides discussion on the experimental results both on Human and Yeast data while conclusions are drawn in Section 5.

2 Review of Existing PPI Inference Models:

This Section will provide an overview of the existing PPI inference models. The models presented here are the only ones which combine different biological data to infer PPI. For detailed overview on PPI inference methods we refer interested readers to [1-3].

2.1.1 Similarity based Network Inference [4]:

In this approach kernel K is used as a similarity metric. If the kernel value $>$ threshold Γ then proteins are considered to be interacting and edge is drawn between them.

2.1.2 Kernel Canonical Correlation Analysis (KCCA)

In this approach the vertices data is mapped to the Euclidean space then similarity based approach (Section 2.1.1) is used to find edges between them. To map the input to the Euclidean space, the known network (partial interaction network or training data) is transformed to a positive semi definite matrix. For this, usually diffusion kernel is used [5]. The correlating directions are then searched in the transformed space and the geometric data by *Canonical Correlation Analysis* . The parameters for the model are canonical directions d, and a regularization parameter λ. To compare fairly with the model we used the $d = 20$ and $\lambda = 0.01$ as the model was reported to have best accuracy for these parameters by Yamanishi *et al* [4].

2.1.3 Kernel Metric Learning (KML)

In this method a metric is learned from the training data before apply approach mentioned in Section 2.1.1 to bring connected vertices closer and non-connected vertices to be further apart from each other [1]. The KML algorithm also like KCCA depends on a regularization parameter λ and a dimension of projection. Vert *et al*; [6] observed best results for $d = 20$ and $\lambda = 2$ and therefore we've used the same parameters in the report.

2.1.4 Matrix Completion with EM (MCEM):

The MCEM algorithm fills missing entries in the adjacency matrix by using kernel matrix obtained from the genomic data. The entries are filled to minimize the geometric distance with the resulting complete matrix. The method is parameter free and has closed form solution [1].

3 Hyper Scalable Network Inference Machine:

The hSNIM model consists of four layers (Figure 2). The layers are:

1- Pre-Processing (Layer 1)

2- Individual Model Reconstruction (Layer 2)

3- Networks Integration (Layer 3)

4- Meta Learning (Layer 4)

Each of the layers and how they are combined will be described in the following subsections.

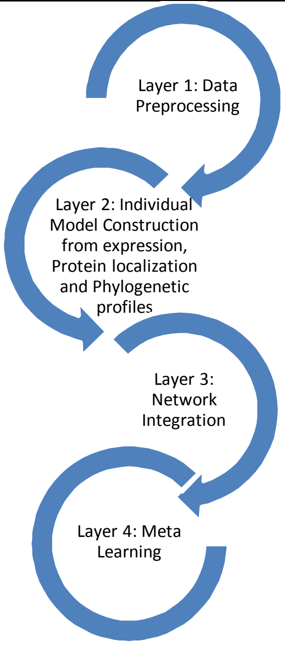
Layer 1: Data Preprocessing
Layer 2: Individual Model Construction from expression, Protein localization and Phylogenetic profiles
Layer 3: Network Integration
Layer 4: Meta Learning

Figure 2. Hyper Scalable Network Inference Machine

3.1.1 Pre-Processing

The hSNIM first pre-processes the data and adds additional features to improve the prediction accuracy. The data input set *X* consists of expression data $\ell \in \mathbb{R}^{m \times n1}$, SCL localization data $\mathcal{L} \in \mathbb{R}^{m \times n2}$, phylogenetic profiles $\mathfrak{I} \in \mathbb{R}^{m \times n3}$, and partial PPI networks $\mathscr{g}$.

The expression data ℓ is first normalized to have mean zero and $\sigma^2 = 1$. Then for each pair of protein $\{\pi_j, \pi_k\}$, spearman correlation ρ is calculated. It adds additional information of gene co-regulation. After this data is discretised using Fayyad *et al;*[7] algorithm. The rationale behind this was that we wanted to test the accuracy of the models against Naïve Bayes. However, it is an optional step and can be ignored since if decision trees are used, as in hSNIM, they are capable of dealing with the continuous values. The expression pre-processed data $\hat{\ell} \in \mathbb{R}^{m \times n1+1} = [\ell, \rho]$ is then generated by concatenating $\ell^j, \ell^k, and\ \rho.$

The SCL localization data $\mathcal{L} \in \mathbb{R}^{m \times n2}$is a binary valued data where

$$\left.\begin{array}{r}\mathcal{L}_i = 1; \text{If protein in located in the cellular compartment } i \\ \mathcal{L}_i = 0; \text{If proteins is not located in the cellular compartment } i\end{array}\right\}$$

For a given pair of proteins $\{\pi_j, \pi_k\}$, hSNIM calculates conditional probabilities $\mathrm{P}(\mathfrak{I}_i^j = \mathfrak{I}_i^k)$ in the pre-processing steps and adds this P-value to the data. The rationale behind this step is that if the proteins are co-located in a compartment then there is a high chance that they'll interaction. However, if the proteins are never collocated in a compartment they'll not have physical interaction. The SCL pre-processed data $\hat{\mathcal{L}} \in \mathbb{R}^{m\times n2+1} = [\mathcal{L}, \mathrm{P}]$ is then generated by concatenating $\mathcal{L}^j, \mathcal{L}^k, and$ P.

Similarly, the phylogenetics data $\Im \in \mathbb{R}^{m\times n3}$is a binary valued data where

$$\left.\begin{array}{r}\Im_i = 1; \text{If protein is conserved in species } i\\ \Im_i = 0; \text{If protein is not conserved (or absent) in species } i\end{array}\right\}$$

Like SCL, for a given pair of proteins $\pi = \{\pi_j, \ \pi_k\}$, hSNIM calculates conditional probabilities $\mathrm{P}(\Im_i^j = \Im_i^k)$. This provides additional information that if both proteins are conserved in the same species then they can probably interact. The phylogenetic pre-processed data $\widehat{\Im} \in \mathbb{R}^{m\times n3+1} = [\Im, \mathrm{P}]$ is then generated by concatenating $\Im^j, \Im^k, and$ P.

3.1.2 Individual Model Construction

In the next step the hSNIM generates individual models $h_\ell(.), h_\mathcal{L}(.)\ and\ \ h_\Im(.)$ By training classifiers on $\hat{\ell} \in \mathbb{R}^{m\times n1+1}$ $\hat{\mathcal{L}} \in \mathbb{R}^{m\times n2+1}$ and $\widehat{\Im} \in \mathbb{R}^{m\times n3+1}$. Each model $h_\ell(.), h_\mathcal{L}(.)\ and\ \ h_\Im(.)$ is trained by its corresponding $\hat{\ell}$, $\hat{\mathcal{L}}$ and $\widehat{\Im}$ data. For a given example protein pair π^i each of the models then generates individual output $\widehat{y_\ell}^i, \widehat{y_\mathcal{L}}^i\ and\ \widehat{y_\Im}^{\ i}$. These outputs are then concatenated to form $\hat{Y}^i\ = \{\ \widehat{y_\ell}^i, \widehat{y_\mathcal{L}}^i\ and\ \widehat{y_\Im}^{\ i}\}$. The hSNIM used C4.5 decision trees [8] to generate the hypothesis $h_\ell(.), h_\mathcal{L}(.)\ and\ \ h_\Im(.)$. We've used standard entropy as cut-off criteria in these decision trees. For details we refer interested readers to [9].

3.1.3 Integrated Model

Once the intermediate output $\hat{Y}^i$ is generated, in this step, the final interaction probability of interaction is computed by logistic regression using:

$$\log\frac{P(y=1\,|\,x)}{P(y=0\,|\,x)} = \alpha + \beta.x^T \tag{1}$$

where $P(y=0\,|\,x) = \dfrac{1}{1+e^{\alpha+\beta.x^T}}$ and $P(y=1\,|\,x) = \dfrac{e^{\alpha+\beta.x^T}}{1+e^{\alpha+\beta.x^T}}$

In this model Parameter α and β are computed using Maximum likelihood.

3.1.4 Error Reduction through Meta Learning

The hSNIM uses bagging by Breiman [10] in the last stage. The choice of bagging over boosting here is empirical where we observed improved results when trees were combined using bagging.

4 Results and Discussion

To rigorously test the performance of the models we used Yeast data from (<et al). The data contains interactions of 668 × 668 proteins. Missing values were removed from the data so that models could be compared. The data set X consists of expression data $\ell \in \mathbb{R}^{668\times 157}$, protein localization data $\mathcal{L} \in \mathbb{R}^{668\times 22}$, phylogenetic profiles $\Im \in \mathbb{R}^{668\times 145}$, and partial PPI networks $\mathcal{G}$. For the sake of fair comparison for Yeast data we only kept high confidence interactions

supported by several experiments which, after removal of proteins without interaction, resulted in a total of 2782 edges amongst 668 proteins as used by Yamanishi *et al*, [4]. The negative data was randomly generated and to get confidence in the experimental results we used 10 fold cross validation and ran models 1000 times to collect the AUC. The hSNIM model is tested against KCCA, KML and MCEM. For KCCA and KML we've used the parameters mentioned in Section (2).

Table 1 shows average AUC of 1000 runs for each model. The hSNIM clearly performed better than the comparative models for the yeast data set. Figure 3, shows example ROC plots generated when hSNIM was used to infer Yeast PPI network.

Table 1. Area under the Curve (AUC) for the Yeast data

Model	**hSNIM**	**KML**	**KCCA**	**MCEM**
AUC	**0.82**	0.77	0.74	0.79

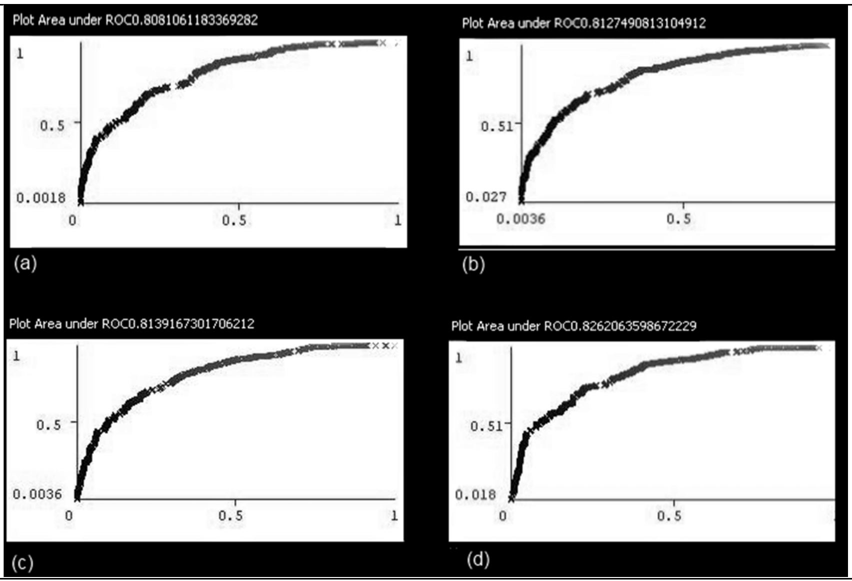

Figure 3. ROC plots for the Yeast PPI network

To further validate the performance of hSNIM we used tested the model to infer human interactome. Inference of human PPI network poses new challenges since the data is not only high dimensional but also there is very little known about human PPI network and therefore, there are lots of missing values unlike Yeast.

Table 2 shows accuracy of individual classifiers in Layer 2 (Figure 2). We used Naïve Bayes as a baseline for the comparison and as expected Decision trees performed better than Naïve Bayes models. A clear improvement can be observed when only Phylogenetic data was used to infer PPI. Decision trees showed 78.10% accuracy versus 56.83% in this case and thus were a clear choice to be part of the model.

Table 3 shows when bagging was applied to the models in Layer 2 and it can be observed that bagging significantly improved the prediction accuracy. Again clear improvement was observed in the case of phylogenetic data where bagging improved accuracy from 78.10% t0 83.31%. However, the most improvement is seen when hSNIM (Integrated) was used to infer human PPI network.

Tables 2 and 3 also show that the most important PPI predictor data in this case are phylogenetic profiles while the weakest predictor data are SCL. The reason for SCL being a weak predictor is also because for human the data is far from complete and most of the proteins don't have localization information attached to them. We also did test hSNIM when we included all the proteins which have at least one of the three dataset available and the average accuracy dropped to 71%.

It should be noted that we've equal number of positive and negative examples in the human test data and therefore accuracy is a reasonable metric here.

Table 2. Accuracy of the models on Human Data

Data Types		**Phylogenetics**	**Expression**	**SCL**
Models	**Naïve Bayes**	56.83	58.70	54.68
	Decision Trees	78.10	63.41	56.64

Table 3. Accuracy of the models with Bagging on Human Data

Data type	**Phylogenetics**	**Expression**	**SCL**	**hSNIM (Integrated)**
Accuracy	83.31	67.29	56.64	**98.79 (± 0.002)**

Tables 3 also underpins that the best case scenario is when all the data is present for prediction, in which case accuracy is 98.79 but in the worst case scenario when only SCL data is present it can dramatically drop to 56.67%. It is important to note, however, that expression data is collected using high throughput microarrays so the chances to have expression data missing for the coding gene are rare. The comparative models couldn't be tested for human data since the current implementation for them required complete data, however the results on the Yeast data show that hSNIM has shown better performance than KML, KCCA and MCEM and also has the ability to infer networks in the presence of missing data.

5 Conclusions:

This report presented *Hyper Scalable Network Inference Machine* (hSNIM) to infer Protein-Protein Interaction (PPI) networks. The model not only scales to infer human interactome (human PPI network) and has unique ability to deal with missing information but also can provide the best accuracy compared to the existing methods. The model achieved this accuracy by the virtue of fusing various forms of data like Gene Expression, Partial PPI data, Sub-cellular localization and Phylogenetic tree. The experimental results underpin that hSNIM has highest AUC 0.82 compared to existing proposed models like Kernel Metric Learning (KML), Kernel Canonical Correlation Analysis (KCCA) and Matrix Completion with EM (MCEM). The model also achieved 98.79% accuracy when it was applied to infer human interactome, which underscores its ability to not only infer large scale PPI networks but also demonstrates that it can achieve low False Positive Rate.

References:

[1] K. Bleakley, *et al.*, "Supervised reconstruction of biological networks with local models," *Bioinformatics,* vol. 23, pp. i57-65, Jul 1 2007.

[2] T. Berggard, *et al.*, "Methods for the detection and analysis of protein-protein interactions," *Proteomics,* vol. 7, pp. 2833-42, Aug 2007.

[3] E. M. Phizicky and S. Fields, "Protein-protein interactions: methods for detection and analysis," *Microbiol Rev,* vol. 59, pp. 94-123, Mar 1995.

[4] Y. Yamanishi, *et al.*, "Protein network inference from multiple genomic data: a supervised approach," *Bioinformatics,* vol. 20 Suppl 1, pp. i363-70, Aug 4 2004.

[5] R. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," *ICML,* 2002.

[6] J.-P. Vert and Y. Yamanishi, "Supervised graph inference," *Advances in Neural Information Processing Systems,* 2005.

[7] U. M. Fayyad and K. B. Irani, "On the Handling of Continuous-Valued Attributes in Decision Tree Generation," *Machine Learning,* vol. 8, pp. 87-102, 2002.

[8] J. R. Quinlan, "Improved use of Continuous Attributes in C4.5," *Journal of Artificial Intelligence Research,* vol. 4, pp. 77-90, 1996.

[9] J. R. Quinlan, *C4.5: Programs for Machine Learning*: Morgan Kaufmann Publishers,, 1993.

[10] L. Breiman, "Bagging Predictors," *Machine Learning,* vol. 24, pp. 123–140, 1996.