REFINING LITERATURE CURATED PROTEIN INTERACTIONS USING EXPERT OPINIONS

OZNUR TASTAN*[,1], YANJUN QI[2], JAIME G. CARBONELL[3], JUDITH KLEIN-SEETHARAMAN[4]

[1] *Department of Computer Engineering, Bilkent University, Cankaya, Ankara, Turkey*
[2] *Department of Computer Science, University of Virginia, Charlottesville, VA, USA*
[3] *Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA*
[4] *Division of Metabolic and Vascular Health, University of Warwick, Warwick, Coventry, UK*
**E-mail: oznur.tastan@cs.bilkent.edu.tr*

The availability of high-quality physical interaction datasets is a prerequisite for system-level analysis of interactomes and supervised models to predict protein-protein interactions (PPIs). One source is literature-curated PPI databases in which pairwise associations of proteins published in the scientific literature are deposited. However, PPIs may not be clearly labelled as physical interactions affecting the quality of the entire dataset. In order to obtain a high-quality gold standard dataset for PPIs between human immunodeficiency virus (HIV-1) and its human host, we adopted a crowd-sourcing approach. We collected expert opinions and utilized an expectation-maximization based approach to estimate expert labeling quality. These estimates are used to infer the probability of a reported PPI actually being a direct physical interaction given the set of expert opinions. The effectiveness of our approach is demonstrated through synthetic data experiments and a high quality physical interaction network between HIV and human proteins is obtained. Since many literature-curated databases suffer from similar challenges, the framework described herein could be utilized in refining other databases. The curated data is available at `http://www.cs.bilkent.edu.tr/~oznur.tastan/supp/psb2015/`

Keywords: Protein-protein Interactions, Literature Curated Databases, Crowd-Sourcing

1. Introduction

Literature curated databases[1–8] extract PPIs from published articles, organize them and make them available online. In addition to supplying prior knowledge for future experiments on a specific protein function, these sets of interactions enable system level analyses of interactomes, serve as benchmark data to quantify error rates in high-throughput experimental assays[9] or they are used as training/testing data to build predictive models.[10] Such analyses depend critically on the inherent quality of the data.

Obtaining a high quality set of direct physical PPI interactions often presents a challenge. In some cases, databases include a mixture of functional indirect associations and direct physical interactions, without specifying the distinction. For example, the work presented here is motivated by our previous study on predicting the HIV-1-human interactome,[11] where we faced the challenge of extracting the subset of direct physical interactions from the NIAID HIV-1, Human Interactions database,[2,3] which does not readily provide a reliable direct physical interaction set. Since this is a general problem, the International Molecular Exchange (IMEx) consortium[12–14] has announced that literature-curated databases should provide more detailed information about the experiments using structured ontologies describing the type of the interaction or the experimental techniques employed. In theory, these experimental details might allow the user or the database curator to review each interaction and use their own judgment to decide how reliable a pair is. In practice, this route is time consuming and is not

guaranteed to arrive at good quality datasets. Users either assume all small-scale experiments are of equally high quality[15] or disregard some portion of the interactions based on additional criteria such as the type of experimental technique, or the number of publications that validates the interaction. Few databases provide reliability scores, e.g. the Molecular Interaction Database (MINT)[16,17] that combines information such as the scale of the experiment, the type of the experiment, the number of publications supporting an interaction, and the presence or absence of ortholog interactions. However, the scoring function blends in several parameters to quantify reliability.[16,17]

Assessing whether there is enough evidence to conclude that a reported association is a direct physical interaction requires a complex judgment. One has to concurrently account for the methods employed, the proteins under study, and the results of each specific study. Some experimental techniques are more conclusive than others and techniques do not work uniformly well across all proteins. In addition to the variability in the power and limitations of each technique, the conditions under which a study is conducted are important: *in vitro* or *in vivo* environment, the strains used, the nature and positions of mutations or labels introduced. All such parameters should be taken into account when interpreting the results. Such a complex judgment can be best provided by domain experts.

Although domain experts can provide high quality labels for PPIs, different experts might have different opinions, especially when there is not enough evidence accumulated in the literature to give a perfectly conclusive answer. Additionally, disagreements among experts arise because of their biases, expertise and/or stringency levels; e.g., some experts are more difficult to convince with partial evidence or results of certain experimental techniques than others. Despite these limitations it has been demonstrated in several other domains, that harnessing the power of human judgments collectively – although imperfect individually – can be invaluable for solving difficult tasks.[18]

A long line of work exists in the biostatistics and epidemiology literature where latent variable models have been used to estimate the observer error rates based on results from multiple diagnostic tests without a ground-truth set.[19] Recently, "learning from crowds" has become a very active research topic in machine learning and has already had several successful applications.[18,20,21] For example, through the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project, Marbach et al.[22] used community-based methods to construct high-confidence gene regulatory networks for E. coli and S. aureus. Their integration strategy through "learning from crowds" was shown as a powerful and robust tool for the inference of transcriptional gene regulatory networks. Similarly, Lu et al.[23] combines multiple annotations in the Gene Normalization (GN) challenge in BioCreative III.

In this paper, we adopt a crowd-sourcing strategy to revise an otherwise noisy and heterogeneous literature curated HIV-1, human PPI dataset by collecting expert opinions on PPI. We apply a probabilistic approach to estimate experts labeling accuracies in the absence of a benchmark dataset similar to the approach proposed by Raykar et al.[20] Using these estimations, the probability of the literature curated pairs to reflect direct interactions is assessed, which results in a high quality set of labels for HIV-1, host PPIs. We verify the utility of this approach through synthetic data experiments as well as performance tests conducted with a

new model that is trained with these high quality labels.

2. Literature Curated HIV-1, Human PPIs and Collected Experts' Labels

2.1. *Existing Literature Curated HIV-1, Host PPIs:*

We retrieved literature curated HIV-1 and human PPIs from the NIAID HIV-1, human protein interaction database (henceforth referred to as NIAID database).[2,3] The set includes 2589 HIV-1, human PPIs between 1448 human proteins and 18 HIV-1 proteins. Whether the reported association is a direct physical interaction or not is not provided by the database. Instead, the database describes each interaction by one or more descriptive key phrases such as " interacts with" or "binds", which are extracted from publications reporting these interactions.

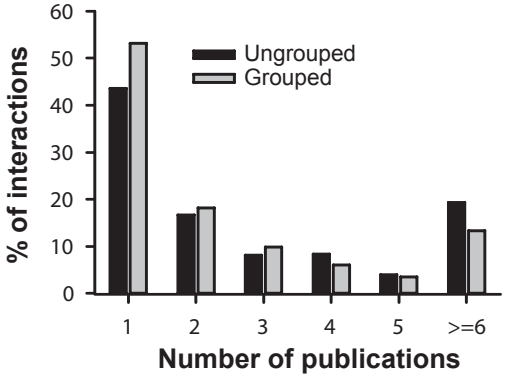

Fig. 1. Distribution of the number of publications supporting each HIV-1, human protein interaction in the NIAID database.[2,3] The graph depicts two analyses, where publications are counted individually (black) or publications are grouped together if they share a common author (gray).

We retrieve the list of publications supporting each interaction in the NIAID database and conduct an analysis on the number of supporting publications for each PPI. Interestingly, 44% of all the PPIs in the database are reported only in a single publication (Fig. 1). When the publications that share at least one common author are grouped together, the statistics become even more striking; the proportion of PPIs supported by a single group is as high as 53% (Fig. 1). The number of PPIs that are supported by more than five publications constitutes only 19 and 13% of all PPIs when publications are ungrouped and grouped, respectively. The lack of follow-up studies by other labs for a given PPI hints at the possibility that for most interactions, there may not be enough experimental validation for their inclusion in a gold standard dataset. This justifies the need to develop a good way to assign confidences to the associated pairs.

2.2. *Collecting Opinions from Experts*

To obtain judgments on interactions published, we contacted a number of experts working in the HIV field. 16 experts provided their opinions on the interactions. One of the experts is a PhD student working with HIV-1 experimentally; all others were professors at different universities who have worked several years on one or more HIV-1 proteins experimentally. The experts were asked to annotate only the interactions of the HIV-1 proteins for which they consider themselves experts. For each HIV-1 protein, an Excel file was prepared, in

which interactions with the HIV-1 protein were listed. The file included the human protein interaction partners of the HIV-1 protein, the keywords retrieved from the NIAID database, and hyperlinks to the original publications so the experts could check the articles if necessary. For HIV-1 proteins, where the number of PPIs are ≤ 50, all the interactions reported in the NIAID database were sent to experts. In cases where > 50 interaction partners were listed for HIV-1 proteins, experts were only provided with the subset of PPIs described with the keywords "interacts with" and "binds". This was to avoid experts feeling overwhelmed with a long list. We adopted this route to increase the chance of receiving a response. In some cases, experts did not use the files; instead, they provided us with a set of interactions which they thought are real and direct PPIs.

3. Model to Estimate Expert Labeling Qualities and PPI Labels

Presented with the same protein pair and the accompanying published evidence, experts sometimes disagreed on whether to label the PPI pair as a direct interaction or not. When there is not enough evidence accumulated in the literature, disagreements among experts might arise because of their biases, expertise, and/or stringency levels. For these reasons, expert opinions will be noisy and subjective. By asking several experts about the same interaction pair, the reliability of the interaction being a direct interaction can be better assessed. Although having as many expert opinions as possible is beneficial, we were not able to obtain multiple expert opinions on all protein pairs due to time or expertise constraints. Thus, there is variance in the number of expert opinions for some pairs.

Taking these considerations into account, the computational problem becomes the following: given noisy opinions and with possibly varying number of judgments for each case, how can we accurately decide if a protein pair is more likely a "direct physical interaction"and what is the uncertainity of the label. A commonly used strategy in this situation is to decide on labels based on a majority vote of the expert labels. However, majority voting assumes that all experts provide opinions of equal accuracy. When there is subjectivity and noise in the opinions, majority voting cannot be expected to perform as desired. In our model, we account for that with a probabilistic latent variable model. In this model experts' labeling accuracies and the probability of the association being a direct interaction are estimated jointly.

3.1. *A Probabilistic Model for Expert Opinions*

Let's consider N literature reported PPI and let $y_i \in \mathcal{Z}$ indicate the true and hidden label for the i^{th} PPI, where $\mathcal{Z} = \{0, 1\}$ ("direct physical interaction" or "not", respectively). The labels y_i are unknown. Instead we have multiple expert labels $\mathbf{y}_i = \{y_i^1, y_i^2, \ldots, y_i^M\}$ provided by M different experts. We introduce a parameter that represents the true unknown labeling accuracy of the expert j for the label type z. We propose to model each expert's labeling accuracy using a biased-coin model,[20] which is,

$$\mathbf{P}\left(y^j = z \,|\, y = z\right) = \theta_z^j \tag{1}$$

where $z \in \mathcal{Z} = \{0, 1\}$. This model assumes that when the hidden true label is one, the expert j flips a coin with bias θ_1^j, meaning that the expert j has a probability of θ_1^j to assign correct label 1 to this instance. When the hidden true label is 0, the expert j flips a coin with bias θ_0^j, meaning that the expert j has a probability of θ_0^j to assign correct label 0 to the

instance. The above formulation can model the situation of different experts having different error rates in their annotations of "direct interaction" and "not direct interaction" label types. This model assumes that the parameter vector $\theta^j = (\theta_0^j, \theta_1^j)$ does not depend on the instance x. Also, another parameter representing the *prior* probability of different labels is $p_i^z = \mathbf{P}(y_i = z)$. A procedure to estimate $\Theta = \{\theta^1, \theta^2, \ldots, \theta^M, p\}$ for all experts is presented in detail throughout the next subsection. Using this model, a *soft probabilistic estimate* of the hidden true label can be calculated as follows:

$$g_i(z \mid \Theta) \equiv \mathbf{P}\left(y_i = z \mid y_i^1, y_i^2, \ldots, y_i^M, \Theta\right) \propto \mathbf{P}\left(y_i^1, y_i^2, \ldots, y_i^M \mid y_i = z, \Theta\right) \times \mathbf{P}(y_i = z \mid \Theta) \tag{2}$$
$$= \prod_{j=1}^{M} \mathbf{P}\left(y_i^j \mid y_i = z, \Theta\right) \times \mathbf{P}(y_i = z \mid \Theta) = p_i^z \times [\theta_z^j]^{h(y_i^j = z)} \times [(1 - \theta_z^j)]^{1 - h(y_i^j = z)}$$

where h is the indicator function. Note that we assume that decisions by the experts are conditionally independent given the true label. Thus, we use the following equation to predict the most probable label for an interaction:

$$\hat{y}_i = \arg\max_{z \in \mathcal{Z}} g_i(z \mid \Theta) \tag{3}$$

The uncertainty of this label based on the above equation is defined as:

$$\hat{u}_i(\hat{y}_i) = 1 - \mathbf{P}(y_i = \hat{y}_i \mid \mathbf{y}_i, \Theta) \tag{4}$$

3.2. *Maximum Likelihood Estimator for Experts' Labeling Qualities*

We estimate the parameters Θ through maximum likelihood estimation (MLE):

$$\hat{\Theta}^{\text{mle}} = \arg\max_{\theta} \mathcal{L}(\mathcal{D} \mid \Theta) \tag{5}$$
$$\text{where } \mathcal{D} = \{(y_i^1, y_i^2, \ldots, y_i^M)\}_{i=1,\ldots,N}$$

Below, we show how to estimate $\hat{\Theta}^{mle}$ for the case where every interaction receives opinions from every expert. Later, we refine the model by relaxing this assumption.

Case 1: Every expert provides labels for every example (global annotation case) The log-likelihood of the observed expert opinions:

$$\mathcal{L}(\mathcal{D} \mid \Theta) = \sum_{i=1}^{N} \log \mathbf{P}(\mathbf{y}_i \mid \Theta) = \sum_{i=1}^{N} \log \sum_{z=0}^{1} \mathbf{P}(\mathbf{y}_i \mid y_i = z, \Theta) \times \mathbf{P}(y_i = z \mid \Theta) \tag{6}$$

The last equation marginalizes over the hidden true label, y_i. We assume decisions by the experts are conditionally independent given the true label:

$$\mathcal{L}(\mathcal{D} \mid \Theta) = \sum_{i=1}^{N} \log \sum_{z=0}^{1} \left(\prod_{j=1}^{M} \mathbf{P}\left(y_i^j \mid y_i = z, \Theta\right) \mathbf{P}(y_i = z \mid \Theta) \right) \tag{7}$$

In Eq. 7, $\mathbf{P}\left(y_i^j \mid y_i = z, \Theta\right)$ is the probability of observing expert label $y_{i,j}$ for interaction i, given the true label of that interaction is $y_i = z$ (similar to calculations in Eq. 2):

$$\mathbf{P}\left(y_i^j \mid y_i = z, \Theta\right) = [\theta_z^j]^{h(y_i^j = z)} \times [(1 - \theta_z^j)]^{1 - h(y_i^j = z)} \tag{8}$$

where h is the indicator function. $\mathbf{P}(y_i = z) = p_z$ is the prior probability of a potential pair belonging to class z; it is assumed that this prior probability is the same for all $i = 1 \ldots N$. In

order to estimate Θ, first Eq. 8 is inserted into Eq. 7 and next, the expectation-maximization (EM) algorithm[24] is applied to maximize a lower bound of this incomplete data likelihood, by considering true label y as the hidden variable:

$$\mathcal{L}(\mathcal{D} \mid \Theta) = \sum_{i=1}^{N} \log \sum_{z=0}^{1} \mathbf{P}\left(\mathbf{y}_i, y_i = z \mid \Theta\right) \geq \sum_{i=1}^{N} \sum_{z=0}^{1} \log \mathbf{P}\left(\mathbf{y}_i, y_i = z \mid \theta\right) \tag{9}$$
$$= \sum_{i=1}^{N} \sum_{z=0}^{1} g_i(z) \log \frac{\mathbf{P}(\mathbf{y}_i, y_i = z \mid \Theta)}{g_i(z)}$$

It is iteratively maximized with respect to the probability distribution $g(z)$ and Θ in the expectation and maximization steps, respectively. The derived update equations for step $t+1$ are as follows:

E-step:

$$g_i^{(t+1)}(z') = \mathbf{P}\left(y_i = z' \mid \mathbf{y}_i, \Theta^{(t)}\right) = \frac{\prod_{j=1}^{M} \mathbf{P}\left(y_i^j \mid y_i = z', \Theta^{(t)}\right) \mathbf{P}\left(y_i = z' \mid \Theta^{(t)}\right)}{\sum_{z=0}^{1} \mathbf{P}\left(y_i = z \mid \Theta^{(t)}\right) \prod_{j=1}^{M} \mathbf{P}\left(y_i^j \mid y_i = z, \Theta^{(t)}\right)} \tag{10}$$

M-step:

$$[\theta_{z'}^{j}]^{(t+1)} = \frac{\sum_{i=1}^{N} g_i^{(t)}(z') \times h(y_i^{j'} = z')}{\sum_{i=1}^{n} g_i^{(t)}(z')} \tag{11}$$

The prior probability p^z is an estimate of the class distribution, is derived from majority vote labels based on the MLE solution. The above procedure is repeated until convergence is attained.

Case 2: Experts only provide labels for a subset of examples (subset annotation case) Not every expert might be available to provide labels for each potential interaction due to expertise and time limitations. Thus, the assumption that every expert can report labels for every instance may not hold. In this section, we provide a solution that works when this assumption is relaxed. Let the group of labelers for the interaction i be a subset, $A_i \subset \{1, \ldots, m\}$. It is required that each interaction will have received at least one opinion from at least one expert. In this case, the EM update equations are modified as follows:

E-step:

$$g_i^{(t+1)}(z') = \mathbf{P}\left(y_i = z' \mid \mathbf{y}_i, \Theta^{(t)}\right) = \frac{\prod_{\substack{j=1 \\ j:j \in A_i}}^{M} \mathbf{P}\left(y_i^j \mid y_i = z', \Theta^{(t)}\right) \mathbf{P}\left(y_i = z' \mid \Theta^{(t)}\right)}{\sum_{z=0}^{1} \mathbf{P}\left(y_i = z \mid \Theta^{(t)}\right) \prod_{\substack{j=1 \\ j:j \in A_i}}^{M} \mathbf{P}\left(y_i^j \mid y_i = z, \Theta^{(t)}\right)} \tag{12}$$

M-step:

$$\theta_{z',j'}^{(t+1)} = \frac{\sum_{\substack{i=1 \\ i:j' \in A_i}}^{n} g_i^{(t)}(z') h(y_i^{j'} = z')}{\sum_{\substack{i=1 \\ i:j' \in A_i}}^{n} g_i^{(t)}(z')} \tag{13}$$

Once the expert labeling accuracies are obtained from the above procedure, they can be plugged into Eqs. 3 and 4 to identify the most probable label and the uncertainty associated with it.

4. Synthetic Data Experiments

Since there are no real data with opinions and expert labels, synthetic data experiments were carried out to test the effectiveness of the method. The synthetic data was generated as follows: Given a prior distribution of label types, a set of true labels was first generated randomly. Meanwhile, each expert's true labeling quality on each label type, $\theta_{j,z}$, was assigned uniformly at random. The underlying rationale is that experts are likely to produce better-than-random answers. In the next step, to simulate an expert's opinion on an instance, true labels for data points were randomly converted to incorrect label types with a probability that follows the expert's error rate (1-$\theta_{j,z}$).

Two scenarios were considered: i) in the *global annotation* scenario, every expert produces labels for every example and ii) in the *subset annotation* scenario, each instance receives a label from a subset of labelers (see above). To realize the second scenario , a probability, $\gamma_{j,z}$, is assigned to each expert and label type which defines how often the expert provides labels for label type z. Each $\gamma_{j,z}$ is drawn uniformly at random from the interval [0,1]. $\gamma_{j,z} = 1$ indicates expert j labels all instances contained in class z and $\gamma_{j,z} = 0$ indicates the expert never labels that label type.

4.1. *Baseline Estimators*

The most probable label estimation is compared to the four following estimators; each labels the interaction as a 'direct interaction' if:

(1) the majority of the experts label them as "direct interaction" (*Majority voting*).
(2) there is at least one expert that thinks it is a "direct interaction" (*Single voting*)
(3) there is at least two experts voting for "direct interaction" (*Double voting*)
(4) all the experts agree on the "direct interaction" label (*All voting*)

4.2. *Evaluation of Synthetic Data Experiments*

The synthetic experiments were repeated $n = 300$ times. To measure how accurately the maximum likelihood estimator can recover the true expert labeling accuracies and uncertainties, the average mean squared error (AMSE) was calculated:

$$AMSE(\hat{\theta}^{\text{mle}}) = \frac{1}{n}\frac{1}{2m}\sum_{z=1}^{2}\sum_{j=1}^{m}(\hat{\theta}_{z,j}^{\text{mle}} - \theta_{z,j})^2, \quad AMSE(\hat{u}) = \frac{1}{n}\left(\hat{u}_i(\hat{y}_i, \hat{\theta}^{\text{mle}}) - u_i(\hat{y}_i, \theta)\right)^2 \tag{14}$$

In order to assess whether the accuracy of the final label is correctly assigned, precision and recall rates are reported. The precision is defined as the fraction of true direct interactions that are identified by the method as "direct interaction". On the other hand, recall is the fraction of the correctly identified "direct interaction" pairs among all the pairs that are direct interactions:

4.3. *Preformance Results*

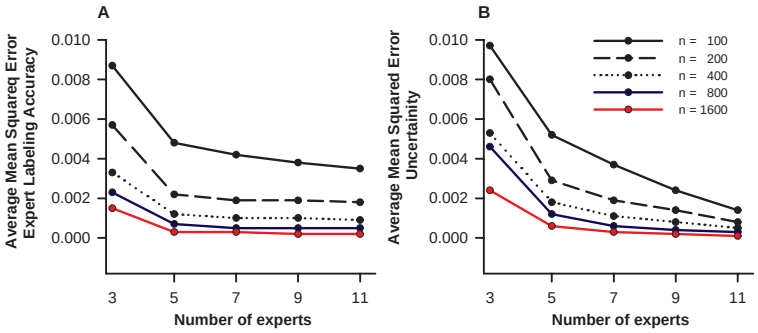

Fig. 2. The average mean squared errors in estimating a) expert labeling qualities and b) uncertainties are plotted as a function of the number of experts for different numbers of pairs to be annotated.

In order to understand the method's robustness for the number of examples and experts that are present, the error rates were measured as a function of the number of experts and number of pairs annotated. Fig. 2 A displays the results for estimated expert label accuracies for the global annotation case. Not surprisingly, the error decreases as more experts are included and more data are provided. Nevertheless, the average MSE of expert labeling accuracies, θ, is $0.0087(\pm 0.0121)$, even for the case with only 3 experts and $n = 100$ data points. Similarly, the error in estimating the uncertainties of data points is not more than 0.010 (see Fig. 2 B). Comparison of error curves for different n reveals that the gain in accuracy decreases in different data regimes. For example, the error in estimating θ decreases by an amount of 0.003 when n is ramped from 100 to 200 and there are 3 experts. This difference is only 0.0008 for cases $n = 800$ and $n = 1600$. A similar trend was observed for the number of experts; the largest gain in accuracy occurs when the number of experts increases from 3 to 5.

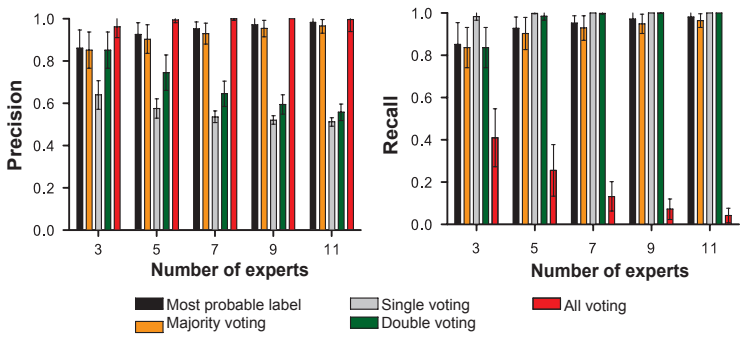

Fig. 3. Precision and recall rates of different labeling strategies.

To assess how well the method can identify true direct interactions, the precision and

recall rates of the estimator are calculated. The precision and recall of the MLE estimator were compared to four other estimators (described in Section 4.1). Fig. 3 displays the precision and recall for different numbers of labelers for the experiments described above and shown in Fig. 2. As can be seen in Fig. 3, single-voting would cover the largest quantity of the true interactions correctly, but it would also consider many incorrect interactions as real physical interactions thus exhibiting a low precision and high recall rate. A similar observation is valid for the double-voting scenario. In both cases the precision becomes worse as the number of labelers increase, since the probability of any pair of labelers producing an incorrect label increases. The opposite trend is true for the all-voting case; the precision is high since the criterion to label a pair as direct interaction is very strict: an agreement between all experts is sought. However, this estimator suffers from low recall rates. In summary, the all-voting strategy results in high confidence sets, but disregards a large portion of the available data; whereas single-voting or double-voting lead to sets with high coverage but both suffer from high false positive rates. The majority-voting method is a robust one; both the precision and recall rates are high, and additionally, as the number of experts increases, the performance improves too. Nevertheless, the maximum likelihood estimator is the best method for both precision and recall rates for all numbers of experts. This is because the noise is taken into account in our probabilistic framework.

5. Refined Literature Curated HIV-1, Human Protein Interactome

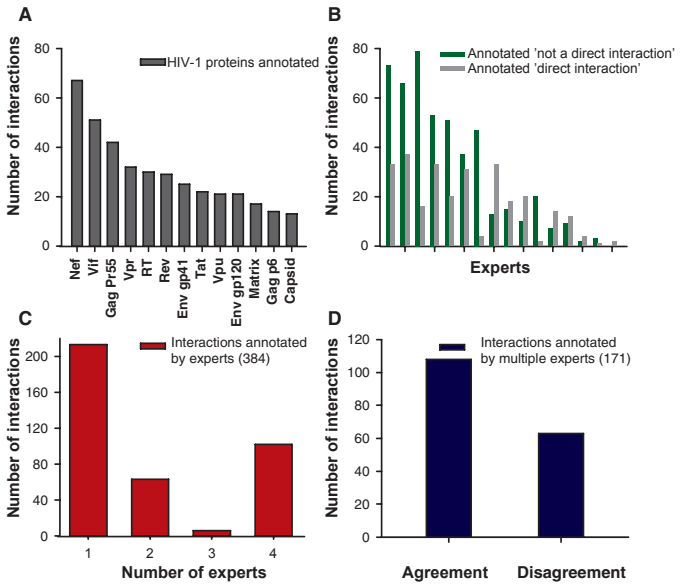

Fig. 4. a) Number of interactions annotated for each HIV-1 protein b) Number of interactions annotated as "direct interaction" (green) and "not a direct interaction" (gray) by each of the 16 HIV-1 experts c) the number of experts annotating each interaction d) Counts of agreed and disagreed interactions when multiple experts assign labels.

In order to estimate expert labeling accuracies, only the interactions that are multiply annotated were considered. For various HIV-1 proteins, different numbers of interactions are annotated; the HIV-1 protein nef has 67 interactions annotated, whereas capsid has only 13 (Fig. 4a). Each HIV-1 expert provided different number of labels (Fig. 4b). The majority of interactions received only one expert opinion (213/384), whereas the rest (171/384) received

multiple expert comments (Fig. 4c). For interactions for which there are multiple opinions from different experts, disagreements among the experts were observed: for 37% of interactions (63/171) experts disagree on the label type (Fig. 4d). Of all the expert annotated interactions, 299 of them are described by the keywords "interacts" with and/or "binds" and these have the most potential to be direct interactions. However, at least two experts still annotated 73 out of 299 as "not a direction interaction" with no disagreements. These results highlight the necessity of reviewing the published interactions with community opinion.

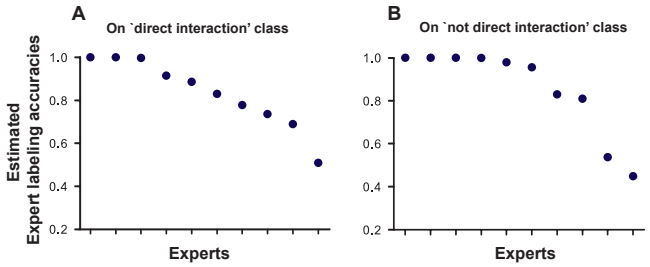

Fig. 5. The estimated labeling accuracies of experts on annotating the PPIs for a) "direct interaction" class and b) "not direct interaction" class.

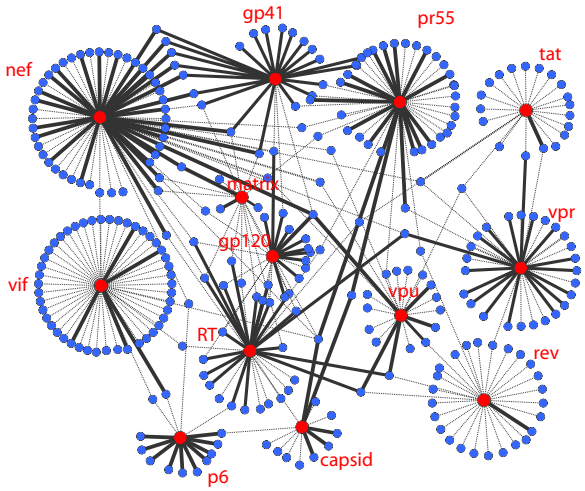

Fig. 6. Refined HIV-1, human protein interaction network based on the estimated confidence scores using HIV-1 expert opinions. Nodes indicate HIV-1 proteins (red) and human proteins (blue). An edge indicates an interaction at least one expert having provided an opinion about it. The thicker the edge, the higher the probability of the pair being a direct interaction according to the estimates. The solid lines represent the interactions for which $\mathbf{P}\,(y = \text{direct interaction}) > 0.5$, whereas dashed lines represent cases for which this probability is < 0.5. The HIV-1 proteins' names are placed next to its subnetwork of interactions.

Using the interactions that receive more than one expert opinion (171 interaction pairs), the experts' labeling accuracies were assessed. There were 16 experts in total, but not all experts provided data for both label types. The estimated label accuracies for the experts are plotted in Fig. 5. 7 out of 10 experts have a labeling accuracy of more than 75 % accuracy on the "direct interaction" class and 8 out of 10 retain this level of accuracy for the "not direct interaction" class. Using the expert labeling accuracies, the most probable class label was calculated for all the annotated interactions. For 147 (out of 384) reported interactions, there was enough evidence to conclude that they have a direct interaction. Fig. 6 displays the resulting network.

6. Validation of Derived Labels with a Supervised Model

Table 1. Performance of the three Random Forest models.

Trained with	**MAP avg**	**MAP stderr**	**AUC avg**	**AUC stderr**
expert positives + expert negatives	0.7144	0.0083	0.7818	0.0052
expert positives + random selected negatives	0.5466	0.0077	0.6392	0.0084
Baseline	0.4298	0.0075	0.4966	0.0098

The estimated labels obtained in this paper provide a high quality set of interaction labels, which include 158 positive examples and 226 negative examples. In the ensuing discussion, we will be designating this set as "expert-labeled". To validate these labels, we explore a supervised prediction strategy. We built models to classify an interaction into two classes of "direct interaction" and "not direct interaction" based on biological features of interacting proteins [a]. A Random Forest classifier was learned using the the expert-curated labels in a cross-validation setting. This model achieves a 71% MAP score, which is significantly better than the baseline whose MAP score is 43% (Table 1). The baseline classifier was trained on the same set of examples while the labels of the training examples were randomly permutated. This indicates that the expert-derived labels correlate with the features more strongly.

The subset of interactions estimated to have the "not direct interaction" label type is especially valuable for the prediction task as there is no negative set of PPI interactions readily available. A common way to circumvent this difficulty is to create negative datasets by selecting randomly paired examples that are not in the positive set.[25] As the randomly selected negative labels are likely to be far away from the class boundary, they are more easily classified and more likely to give optimistic estimates favoring prediction success; that is the decision boundary learned from them might be too far away from the real decision boundary. Conversely, the expert-labeled negatives from our study are more likely to be in close proximity of the class boundary because they are functional associations. Therefore, our expert-labeled negative examples will define a better decision boundary.

In order to judge how our expert-labeled negative data contributes to the supervised prediction model, a third Random Forest classifier[26] was trained. This model uses the expert labeled positive set and a negative label set comprising random pairs that are not reported in the NIAID database. This model performs better than the baseline, 0.55 ($\pm$ 0.0077) (compare to 43%), but it performs worse than the first model which uses the expert labeled negative examples, (compared to 71% MAP). The AUC values are also ranked similarly (See Table 1). All three models were tested on the expert labeled data in a 3-fold cross validation (CV) scenario. The CV procedure is repeated 10 times, where at each repeated run the splitting of the data is different and random. These empirical results strongly indicate that the curated data is highly valuable.

7. Conclusions

In this paper, a set of expert opinions on literature curated HIV-1, host interactions were collected. To account for noise and subjectivity in expert opinions, the expert labeling accuracies

[a]Each PPI pair is described by 42 features derived from diverse set of biological information with details described in our earlier work.[11]

were estimated and these estimates were used to compute reliability scores that rank interactions with their likelihood to be direct physical interactions. A Random Forest model trained with the derived labels validates the quality of the collected data. The scope of the method presented here is not limited to HIV data, but is applicable to other bodies of literature-curated databases, where noisy labels from multiple experts are available and there is no benchmark data to estimate labeler qualities.

Acknowledgement

We are grateful to the 16 HIV-1 experts for sharing their opinions about HIV-1, human dataset and Prof. Ziv-Bar Joseph for valuable discussions. We also thank Pittsburgh Center for HIV Protein Interactions, especially Prof.Chris Aiken and Dr.Teresa Brosenitsch, for expert advice and helping us in reaching out experts. The work has been supported in part by NIH, NIAID P50GM082251, NSF CCF-1144281, NIH-NLM 2RO1LM007994-05, EraSysBio+ grant and BMBF, SHIPREC and EU Marie Curie Actions 626470 MPFP FP7-PEOPLE-2013-IIF. O.T. acknowledges support from Bilim Akademisi - The Science Academy, Turkey under the BAGEP program and the support from L'Oreal-UNESCO under the National Fellowships Programme for Young Women in Life Sciences.

References

1. B. J. Breitkreutz, C. Stark, T. Reguly *et al.*, *Nucleic Acids Res* **36**, D637 (2008).
2. R. G. Ptak, W. Fu, B. E. Sanders-Beer *et al.*, *AIDS Res Hum Retroviruses* **24**, 1497 (2008).
3. W. Fu, B. E. Sanders-Beer, K. S. Katz *et al.*, *Nucleic Acids Res* **37**, D417 (2009).
4. A. Chatr-aryamontri, A. Ceol, L. M. Palazzi *et al.*, *Nucleic Acids Res* **35**, D572 (2007).
5. I. Xenarios, E. Fernandez, L. Salwinski *et al.*, *Nucleic Acids Res* **29**, 239 (2001).
6. S. Kerrien, Y. Alam-Faruque, B. Aranda *et al.*, *Nucleic Acids Res* **35**, D561 (2007).
7. T. S. Keshava Prasad, R. Goel, K. Kandasamy *et al.*, *Nucleic Acids Res* **37**, D767 (2009).
8. G. D. Bader, I. Donaldson, C. Wolting *et al.*, *Nucleic Acids Res* **29**, 242 (2001).
9. H. Yu, P. Braun, M. A. Yildirim *et al.*, *Science* **322**, 104 (2008).
10. B. A. Shoemaker and A. R. Panchenko, *PLoS Comput Biol* **3**, p. e43 (2007).
11. O. Tastan, Y. Qi, J. G. Carbonell *et al.*, *Pac Symp Biocomput* , 516 (2009).
12. S. Orchard, H. Hermjakob and R. Apweiler, *Proteomics* **3**, 1374 (2003).
13. S. Orchard, S. Kerrien, P. Jones *et al.*, *Proteomics* **7 Suppl 1**, 28 (2007).
14. S. Orchard, L. Salwinski, S. Kerrien *et al.*, *Nat Biotechnol* **25**, 894 (2007).
15. Editorial, *Nature Methods* **6**, p. 2 (2009).
16. A. Chatr-Aryamontri, A. Ceol, L. Licata *et al.*, *Trends Biochem Sci* **33**, 241 (2008).
17. L. Licata, L. Briganti, D. Peluso, L. Perfetto *et al.*, *Nucleic Acids Research* **40**, D857 (Jan 2012).
18. L. von Ahn, B. Maurer, C. McMillen *et al.*, *Science* **321**, 1465 (2008).
19. P. Albert and L. Dodd, *Biometrics* **60**, 427 (2004).
20. V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni and L. Moy, *The Journal of Machine Learning Research* **11**, 1297 (2010).
21. P. G. Ipeirotis, F. Provost and J. Wang, Quality management on amazon mechanical turk, in *Proceedings of the ACM SIGKDD workshop on human computation*, 2010.
22. D. Marbach, J. C. Costello, R. Küffner, Vega *et al.*, *Nature methods* **9**, 796 (2012).
23. Z. Lu, H.-Y. Kao, C.-H. Wei *et al.*, *BMC Bioinformatics* **12**, p. S2 (2011).
24. A. Dempster, N. Laird and D. Rubin, *Journal of the Royal Statistical Society, Series B* **39**, p. 138 (1977).
25. A. Ben-Hur and W. S. Noble, *BMC Bioinformatics* **7 Suppl 1**, p. S2 (2006).
26. L. Breiman, *Mach. Learn.* **45**, 5 (October 2001).