Pacific Symposium on Biocomputing 14:516-527 (2009)

# PREDICTION OF INTERACTIONS BETWEEN HIV-1 AND HUMAN PROTEINS BY INFORMATION INTEGRATION

OZNUR TASTAN[1], YANJUN QI[1], JAIME G. CARBONELL[1] AND JUDITH KLEIN-SEETHARAMAN[1, 2†]

*[1]School of Computer Science, Carnegie Mellon University,15213 and [2]Department of Structural Biology, School of Medicine, University of Pittsburgh, 15260, Pittsburgh, PA USA.*

Human immunodeficiency virus-1 (HIV-1) in acquired immune deficiency syndrome (AIDS) relies on human host cell proteins in virtually every aspect of its life cycle. Knowledge of the set of interacting human and viral proteins would greatly contribute to our understanding of the mechanisms of infection and subsequently to the design of new therapeutic approaches. This work is the first attempt to predict the global set of interactions between HIV-1 and human host cellular proteins. We propose a supervised learning framework, where multiple information data sources are utilized, including co-occurrence of functional motifs and their interaction domains and protein classes, gene ontology annotations, posttranslational modifications, tissue distributions and gene expression profiles, topological properties of the human protein in the interaction network and the similarity of HIV-1 proteins to human proteins' known binding partners. We trained and tested a Random Forest (RF) classifier with this extensive feature set. The model's predictions achieved an average Mean Average Precision (MAP) score of 23%. Among the predicted interactions was for example the pair, HIV-1 protein tat and human vitamin D receptor. This interaction had recently been independently validated experimentally. The rank-ordered lists of predicted interacting pairs are a rich source for generating biological hypotheses. Amongst the novel predictions, transcription regulator activity, immune system process and macromolecular complex were the top most significant molecular function, process and cellular compartments, respectively. Supplementary material is available at URL www.cs.cmu.edu/~oznur/hiv/hivPPI.html

## 1. Introduction

### 1.1. *Motivation*

Human immunodeficiency virus-1 (HIV-1) is the etiologic agent of acquired immune deficiency syndrome (AIDS) and continues to be a major health threat [1, 2]. The number of AIDS-related deaths was ~2.1 million in 2007 alone [3];

† Corresponding author

an estimated ~33.2 million people worldwide are infected [3]. HIV-1 contains a single stranded RNA genome, which codes for only 15 proteins; thus, it relies on human cellular functions [4]. The virus exploits the host cell's machinery such that it can successfully produce its progeny while at the same time avoid the immune system. Protein-protein interactions (PPIs) between HIV-1 and its host are vital at every step of the virus life cycle [1]. A recent functional genomic screen using siRNA technology revealed several hundred proteins critical for HIV-1 function [5]. An extensive survey of individual interactions described in the literature [6] has retrieved ~2500 pairs, of which ~1000 are likely direct physical interactions. Here, we propose to use data integration methods on these reported interactions in conjunction with a variety of different biological information sources to predict new PPIs between human and HIV-1.

### 1.2. *Related Work*

In order to identify PPIs in general, many experimental methods are available [7]. Computational methods have been useful in assisting the experimental efforts by either prioritizing PPIs to be tested by subsequent experiments or by validating (or refuting) high-throughput screens [8]. Such computational methods include those based on over-representation of domain or motif pairs observed in interacting protein partners [9-13], or the conservation of gene neighborhood and gene order [14], gene fusion events [15, 16], or the co-evolution of interacting protein pair sequences [17, 18]. Others designed kernels specific for PPIs that make use of the sequence signatures in interacting pairs [19] and protein structural information [20-22]. The large amount and heterogeneity of these multiple indirect and direct information sources suggests to integrate them in a supervised learning framework [23-26].

The computational methods above were applied to predicting PPIs within a single organism ("intra-species prediction"). In contrast, computational work on predicting PPIs between organisms ("inter-species prediction"), especially between host and pathogens has been limited. Dyer *et al.* [27] studied human-*Plasmodium falciparum* interactions and estimated the probability of interaction based solely on sequence signature information [11]. Davis *et al.* [28] studied ten host-pathogen PPIs (not including HIV-1) using structural evidence with a comparative modeling approach: the host, pathogen protein pairs that share similarity to protein complexes with known structures are used to build 3d structural models of putative complexes and pairs with high quality models are filtered by functional and genomic experimental information. The applicability of these methods in the specific case of HIV-1, human prediction task is limited: most of the HIV-1 proteins do not contain domains, statistics on which would be

required in [27] and not all HIV-1 proteins have high similarity to proteins with known structure which is necessary for [28].

### 1.3. *Approach*

Here, we propose a supervised learning framework to predict PPIs between HIV-1 and human proteins by integrating multiple biological information sources. This framework has been found useful in integrating heterogeneous biological information to predict intra-species interactions [23-26]. However, the inter-species prediction task is comparatively more challenging: high-throughput interaction datasets from yeast two hybrid and mass spectrometry based screens are not yet available publicly and many other useful sources such as co-expression pattern of genes [24] are not directly applicable. Nevertheless, there is another potentially exploitable source of information: the human protein interactome (set of intra-human interactions) itself. The proteins that the pathogen will target should in principle depend on interaction relationships between human proteins because the virus makes use of the existing communication pathways within the cell. Capitalizing on this source of information, we developed a number of features incorporating existing knowledge of human intra-PPIs and integrated these features together with other available information sources to predict HIV-1, human interactions.

## 2. Data and Methods

Details can be found in the supplementary online material Methods section.

### 2.1. *Classification Framework*

The task is formulated as a binary classification problem, where each protein pair belongs to either the 'interaction' or 'non-interaction' class. A protein pair is described by a feature vector, where each feature is derived from one or more biological information sources. We solve the classification problem utilizing the Random Forest classifier (RF) [29]. We chose the RF based on its robustness in scenarios where the features are noisy and redundant, as is the case here. The RF method has been highly successful in predicting intra-species PPIs [24, 26]. Initial tests comparing other classifiers for the HIV-1, human inter-species PPI prediction task also suggested the RF to be suited best (data not shown).

### 2.2. *Dataset*

**2.2.1** ***Interacting protein pairs****:* Interactions between HIV-1 and cellular proteins reported in the scientific literature were retrieved from the NIAID database [6]. The dataset included 2512 interactions involving 1406 human

proteins. Note that in addition to the 15 HIV-1 proteins, the database includes interactions for the precursors of the envelope (env gp160) and gag (gag pr55) and the gag product, p1. We excluded gag p1 due to the limited information available for this protein. Each interaction in the database is represented by one or more descriptive key phrases. Some entries are more likely associated with direct physical PPIs than others (e.g. "interacts with" as compared to "causes accumulation of"). Based on the keywords we grouped the interactions into two exclusive groups: Group 1 represents most likely direct physical interactions. In Group 2, interactions may also be indirect (Supplementary Table S1). Group 1 interactions included 1063 protein pairs involving 721 human proteins. Group 2 included 1447 pairs involving 914 human proteins. The Group 1 interactions constituted the "interaction class" and were used in model building and testing, whereas Group 2 was used to mine the final predictions.

**2.2.2** ***Non-interacting protein pairs:*** Since it cannot be proven that two proteins do not interact, there is no "gold standard" negative set available. For training and testing purposes in the PPI prediction task it is therefore common to choose protein pairs uniformly at random from the set of protein pairs which are not known to interact, and treat them as negative interactions [30]. This is rationalized by the fact that the probability that two randomly chosen proteins interact is small and most methods are able to handle contamination from the small number of potential false negative (FN) matches. In the testing phase, the negative to positive ratio was set as 100:1, a value chosen based on the average number of interactions involving HIV-1 proteins. In training, the negative to positive example ratio is optimized for each cross validation step on the training data.

### 2.3. *Features*

We devised a total of 35 features (Supplementary Table S3). Some features are specific to HIV-1, human protein pairs, while others are related only to human or HIV-1 proteins, and some are derived from the human interactome (Fig. 1).

**2.3.1** ***GO similarity features****:* The Gene Ontology (GO) [31] provides a defined vocabulary of protein attributes for molecular function, cellular component and biological process; proteins are annotated with one or more descriptive GO terms. For each of the three ontologies we developed two features: 'pairwise GO similarity' measures the similarity between the HIV-1 and human proteins in a pair, while 'neighbor GO similarity' refers to the similarity between the HIV-1 proteins and the human protein's human interactors (Fig. 1).

**2.3.2** ***Graph properties of the human interactome:*** Three graph property features were derived from topological properties of the human intra-PPI

network: degree, clustering coefficient and betweenness centrality. The degree of a node in a network is the number of its neighbors; whereas clustering coefficient [32] is the ratio of the edges present among its neighbors to the all possible edges that could be present between them. Betweenness centrality for a node is calculated as the fraction of shortest paths between node pairs that pass through the node of interest [33]. A node with a high betweenness centrality is located in a 'bottleneck' and has control over the information flow between other nodes.

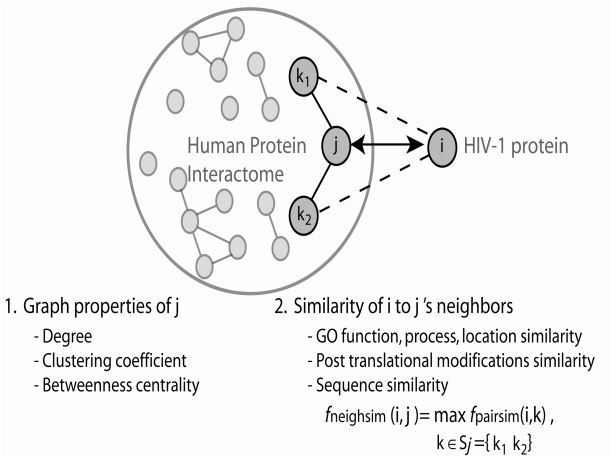


Figure 1. Schematic showing features that incorporate knowledge of the human protein interactome. These features can be conceptually grouped into two categories: 1) graph properties of human protein *j* in human protein interaction network, which include degree, clustering coefficient and betweenness centrality of node *j* 2) the similarity of the HIV-1 protein, *i*, to human protein *j*'s interaction partners denoted by $f_{neigh}$ (i,j) in the figure. The maximal similarity to the neighbors is used. Five features are derived; GO function, process and location similarity in addition to post translational modification and sequence similarity.

**2.3.3** ***ELM-ligand feature:*** Functional sequence motifs which mediate binding were downloaded from the Eukaryotic Linear Motif (ELM) database [34]. For example, the sequence pattern PXXDY (ELM id: LIG_SH3_5) is recognized by SH3 domains. The feature evaluates whether an ELM motif is found in a given HIV-1 sequence and its ligand domain is present in the human protein. The HIV-1 sequences mutate rapidly and some motifs are very short. To avoid false positive matches, we only consider a motif match if it is conserved in multiple HIV-1 sequences; the feature value is weighted with the specificity of the motif.

**2.3.4** ***Gene expression features:*** Four features reflect differential expression patterns of human genes across HIV-1 infected vs. uninfected samples.

**2.3.5** ***Tissue feature:*** This feature encodes whether the tissues that the human proteins are expressed in are susceptible to HIV-1 infection or not.

**2.3.6** ***Sequence similarity features:*** For each pair, two sequence similarity features are employed, pairwise sequence similarity and similarity of the HIV-1 protein sequence to human protein's human binding partners.

**2.3.7** ***Posttranslational modification similarity to neighbor:*** Some PPIs require binding partners to be in a certain posttranslationally modified state and some HIV-1 proteins mimic the posttranslational modification of the human protein's interaction partner [35]. This feature captures if the HIV-1 protein shares any modification with at least one of the interaction partners of the human protein.

**2.3.8** ***HIV-1 protein type features (ptf):*** We included a set of features (one for each HIV-1 protein, with values 1 for pairs that include a particular HIV-1 protein and 0, otherwise) to include the information of how likely an HIV-1 protein is to be in an interaction with a human protein.

### 2.4. *Feature Importance*

In constructing trees in the RF, at each node the attribute causing the highest decrease in the Gini index is chosen as split. Let $p$ denote the fraction of interacting pairs assigned to node $i$ and $1\text{-}p$ the fraction of the non-interacting pairs, the Gini index is computed as $G_i = 2p(1-p)$ [29]. Gini feature importance is derived from the Gini index and is the sum of all decreases in the forest due to a given feature, normalized by the number of trees in the forest.

### 2.5. *Performance Evaluation*

Classifier performance was evaluated with 3-fold cross validation in 10 repeat runs to obtain average values. When evaluating the performance of a classifier on an imbalanced test set such as is the case here, computing accuracy is not useful because a high true-negative (TN) rate can easily be obtained by chance. Therefore, we evaluated the quality of our predictive model using two figures of merits which ignore the success on the TN rate: the receiver operating curve (ROC) and precision vs. recall curve [36]. We employ the Mean Average Precision (MAP) score to summarize the precision vs. recall curve and the area under the ROC (AUC) to summarize the ROC curve as a scalar score which ranges between 0 and 1 [36]. Since the low FP region of the ROC curve is of particular interest in the PPI prediction task, the partial AUC scores R50, R100, R200 and R300 were determined, measuring the area under the ROC curve until reaching 50, 100, 200 and 300 FP predictions, respectively.

### 2.6. *GO Enrichment*

GO enrichment of the human proteins involved in the predicted interactions was identified using Ontologizer 2.0 [37] using the child-term parent intersection method [38] and using Bonferroni correction for multiple hypothesis testing.

## 3. Results and Discussion

### 3.1. *Classifier Performance*

We trained an RF classifier with a rich feature set derived from several biological information sources. The performance of the model was evaluated through 10 repeated 3-fold cross validation experiments. The average precision *vs.* recall curve of these experiments is given in Fig. 2 (solid line). Table 1 lists the average MAP, AUC and partial AUC scores of the model. The model achieves an average MAP score of 0.23 (±0.02) indicating that, on average, of all pairs predicted as interacting, 23% are TP. For PPI predictions, this is a very good performance: because of the highly skewed class distribution the probability of predicting a TP is not 0.5 as when positive and negative pairs are equally distributed, but is ~0.01 (for a ratio positive: negative pairs of 1:100).

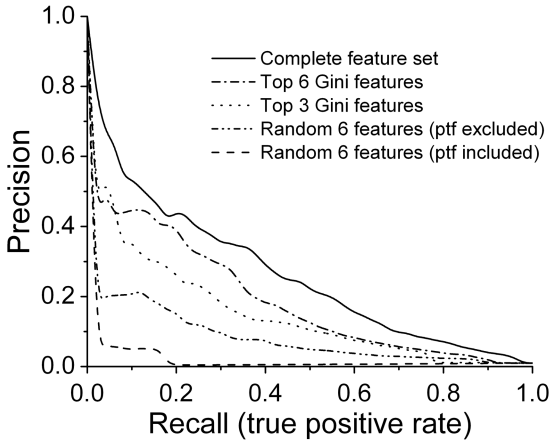


Figure 2. The average precision *vs.* recall curve of the Random Forest model trained on the complete feature set, in comparison to models trained with a subset of features. The top 3 Gini features are degree, betweenness centrality, and GO neighbor process similarity features. The top 6 Gini features are the top 3 Gini features plus clustering coefficient, GO neighbor function, and location features. These are compared to two baseline classifiers, where 6 features are randomly selected from the set of features that does not include the top 6 Gini features, with and without protein type features (ptf).

Table 1. Averages (Avg) and standard deviations (Std) of MAP, AUC and partial AUC scores.

| | MAP | AUC | R50 | R100 | R200 | R300 |
|---|---|---|---|---|---|---|
| **Avg** | 0.2300 | 0.9150 | 0.0670 | 0.1073 | 0.1682 | 0.2156 |
| **Std** | 0.0217 | 0.0120 | 0.0135 | 0.0169 | 0.0204 | 0.0230 |

### 3.2. *Feature Importance*

Biologically, it is of interest to identify the features that contribute the most to the classification of protein pairs. This not only helps reveal relationships between different data sources, but can also suggest which data should be generated by experiments to find novel interactions in this and other host-pathogen systems. We assessed feature importance based on the Gini importance of the RF classifier (see Methods). Strikingly, the graph property and the GO neighbor similarity features are ranked at the top (Fig. 3).

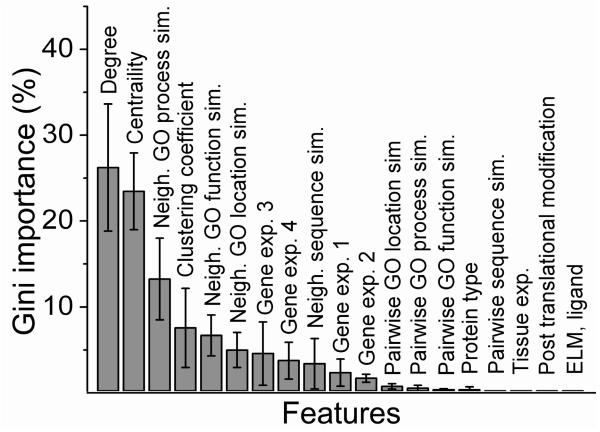


Figure 3. RF Gini importance measures for each feature. Protein type features are grouped together.

To assess the extent to which these features are predictive, we built models using the same train/test data splits as before with only the top 3 and top 6 Gini features (Fig. 2). The top 3 Gini features are degree, betweenness centrality and neighbor GO process similarity (Fig. 3) and the top 6 Gini features in addition include clustering coefficient, neighbor GO function, and cellular location similarities. These models are compared to two baseline models. In the first, the RF classifiers were trained with 6 features selected randomly from the set of features excluding the top 6 Gini features. These random feature sets include the 17 protein type features (ptf), one for each HIV-1 protein. Since these vectors alone do not contain much information, this model forms a weak baseline. A second stronger baseline was built, where the 6 features are randomly selected

from the set of features excluding ptf and the top 6 Gini features. Fig. 2 compares the performance of the above 5 models. The top 6 Gini model performs quite strongly compared to both baselines. However, this model is not as good as the model built using the complete feature set The top 3 Gini model performs significantly worse than the top 6 Gini model, but significantly better than the two baseline models suggesting that the additional top 3 features contain independent and complementary information. Statistical significance of these differences (Fig. 2) was confirmed based on paired t-test comparison of the 30 experiments' MAP scores (at a significance level of 0.05).

The above analysis reveals that the graph features and neighbor similarity features are very informative confirming our intuition in incorporating human interactome knowledge into the model. Graph properties have also been found previously useful in the intra-PPI network prediction task [39-42]. Furthermore, it has been proposed earlier that pathogens exploit network properties of the human interactome: it was shown that the Epstein–Barr virus targets high degree human proteins [43] and it was found recently that pathogens tend to interact with host proteins with high degrees and betweenness centrality [44].

The significant performance difference between the top 6 Gini model and the complete model (Fig. 2) indicates that the low ranked features also contribute to the final performance. For example, the removal of protein type features levels off the precision *vs.* recall curve with respect to the complete feature set (Supplementary Fig. S2). The reason why some of these features' Gini importance scores are very low could be due to their low coverage (Supplementary Table S3).

### 3.3. *Mining Predicted Pairs with siRNA and in Virion Data*

A final model was trained with all available Group 1 interactions, according to the standard methodology that typically the larger the training data, the better the model. We then ranked all HIV-1, human pairs according to their RF score. The score measures the difference between positive and negative votes from the decision trees in the trained RF model and reflects the likelihood of an interaction. The derived ranked order list is available in the supplementary online material. The set of predicted interactions depends on the chosen RF score threshold; lowering the threshold will increase the TP rate at the expense of a higher FP rate. Table 2 (top) presents the number of predicted interactions for different cutoff values. At the lowest threshold we considered (0), 2100 novel interactions are predicted, of which 1 in 5 interactions is expected to be true based on precision measured on the hold out set.

The unknown predicted pairs (Group 2 and novel) were examined in light of the 281 human genes that have been reported in the siRNA screen to have an

effect on HIV-1 infection upon silencing [5] and 314 human proteins highjacked by HIV-1 in its virion [45]. Table 2 (bottom) gives the size of the overlap of our Group 2 and novel predictions with these two datasets. Although the comparison cannot provide means to verify the predictions; the overlapping pairs would be of interest to HIV-1 virologists: the siRNA data provide experimental evidence pointing at their functional relevance and the in virion overlapping set could help differentiate between mere by-stander human in virion proteins from those with functional roles for the virus.

Table 2. Number of predicted pairs at different choices of RF score cutoff. Average recall and precision was calculated on the held-out test sets in cross-validation experiments. The second part of the table presents the overlap (the number of the predicted pairs including the reported human gene) between the new predictions (Group 2 and Novel) and siRNA [5] and in Virion [45] datasets (for details, see text). 'Interactor' refers to the predicted interactions, where the human protein is at least one of the siRNA reported human protein's interaction partner.

| Cutoff | Predictions | Group 1 | Group 2 | Novel | Recall | Precision |
|---|---|---|---|---|---|---|
| ≥ 0.00 | 3372 | 1040 | 232 | 2100 | 0.51 | 0.20 |
| ≥ 0.50 | 1942 | 1034 | 141 | 767 | 0.37 | 0.29 |
| ≥ 1.00 | 1440 | 1023 | 68 | 349 | 0.26 | 0.36 |
| ≥ 1.50 | 1085 | 894 | 34 | 157 | 0.18 | 0.41 |
| ≥ 2.00 | 622 | 538 | 15 | 69 | 0.13 | 0.47 |
| ≥ 2.50 | 279 | 243 | 8 | 28 | 0.09 | 0.47 |
| | Group2 | | | Novel | | |
| | in Virion | siRNA | Interactor | in Virion | siRNA | Interactor |
| ≥ 0.00 | 34 | 4 | 120 | 246 | 46 | 1064 |
| ≥ 0.50 | 24 | 3 | 83 | 101 | 13 | 441 |
| ≥ 1.00 | 10 | 1 | 43 | 48 | 5 | 212 |
| ≥ 1.50 | 5 | 1 | 24 | 17 | 2 | 99 |
| ≥ 2.00 | 3 | 1 | 13 | 8 | 1 | 49 |
| ≥ 2.50 | 2 | 0 | 7 | 4 | 0 | 25 |

### 3.4. *Functions of Predicted Interacting Pairs*

The rank-ordered lists of predicted interacting pairs are a rich source for generating biological hypotheses that can be experimentally validated. For example, we predict that the HIV-1 protein tat and the vitamin D receptor (VDR) interact with a high RF score of 1.96. Tat is a regulatory protein of HIV-1, its main role is to transactivate HIV-1 transcription from the viral long-terminal repeat (LTR) promoter [47]. VDRs are members of nuclear receptors, which act as ligand-inducible transcription factors in response to hormones [48]. Variations at the VDR locus are associated with susceptibility and progression

of AIDS and other immune diseases [49, 50]. The only interaction reported in the NIAID for VDR is a Group 2 interaction with env gp120. Recently, it was reported that tat acts with VDR in a synergistic manner as a stimulator for HIV-1 LTR activity, providing an independent validation of the functional significance of our prediction [51].

A global analysis of the predicted interactions by assessing the enrichment of GO functional terms in predicted Group 2 or novel interactions revealed 31 molecular processes, 19 biological functions and 14 cellular components (Supplementary Tables S5-S7) at significance level 0.01. For example, transcription regulator-, ligand-dependent nuclear receptor-, MHC class I receptor-, and protein kinase C activities are highly enriched molecular functions, while immune system process and response to stimulus are highly represented processes and macromolecular complex, membrane-enclosed lumen and plasma membrane were the top most significant cellular compartments.

## 4. Conclusions

Computational methods can be very effective in assisting experimental efforts in identifying interacting protein pairs within a single organism. This paper extends these methods to predicting interactions between hosts and pathogens, here human and HIV-1. Features derived from multiple genomic and functional data sources and exploiting our knowledge of the human protein interactome were integrated in a supervised learning framework. We inspected our predictions with independent biological datasets to identify the most promising interacting pairs. By providing these new testable hypotheses we hope that our predictions will accelerate experimental efforts to define a reliable network of HIV-1, human protein interactions.

## Acknowledgments

This work is supported in part by National Institutes of Health grants P50-GM082251 and LM07994-01 and National Science Foundation grants EIA0225656/0225636, CAREER CC044917, and the PA Department of Health.

## References

1. A. Trkola, *Curr Opin Microbiol* **7**, 555-9 (2004).
2. D.D. Ho and P.D. Bieniasz, *Cell* **133**, 561-5 (2008).
3. UNAIDS. *2007 AIDS epidemic update.* (2007).
4. A.D. Frankel and J.A. Young, *Annu Rev Biochem* **67**, 1-25 (1998).
5. A.L. Brass *et al*, *Science* **319,** 921-6 (2008).
6. *NIAID HIV-1, Human Protein Interaction Database*. [http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/].

7. B.A. Shoemaker and A.R. Panchenko, *PLoS Comput Biol* **3**, e42 (2007).
8. B.A. Shoemaker and A.R. Panchenko, *PLoS Comput Biol* **3**, e43 (2007).
9. Y. Liu, N. Liu and H. Zhao, *Bioinformatics* **21**, 3279-85 (2005).
10. W.K. Kim, J. Park and J.K. Suh, *Genome Inform* **13**, 42-50 (2002).
11. E. Sprinzak and H. Margalit, *J Mol Biol* **311**, 681-92 (2001).
12. S.M. Gomez, W.S. Noble and A. Rzhetsky, *Bioinformatics* **19**, 1875-81 (2003).
13. T.M. Nye *et al.*, *Bioinformatics* **21**, 993-1001 (2005).
14. T. Dandekar *et al.*, *Trends Biochem Sci* **23**, 324-8 (1998).
15. E.M. Marcotte *et al.*, *Science* **285**, 751-3 (1999).
16. A.J. Enright *et al.*, *Nature* **402**, 86-90 (1999).
17. F. Pazos and A. Valencia, *Protein Eng* **14**, 609-14 (2001).
18. C.S. Goh *et al.*, *J Mol Biol* **299**, 283-93 (2000).
19. S. Martin, D. Roe and J.L. Faulon, *Bioinformatics* **21**, 218-26 (2005).
20. P. Aloy and R.B. Russell, *Bioinformatics* **19**, 161-2 (2003).
21. R. Singh, J. Xu and B. Berger, *Pac Symp Biocomput*, 403-14 (2006).
22. L. Lu, H. Lu and J. Skolnick, *Proteins* **49**, 350-64 (2002).
23. R. Jansen *et al.*, *Science* **302**, 449-53 (2003).
24. Y. Qi, Z. Bar-Joseph and J. Klein-Seetharaman, *Proteins* **63**, 490-500 (2006).
25. X.W. Chen and M. Liu, *Bioinformatics* **21**, 4394-400 (2005).
26. N. Lin *et al.*, *BMC Bioinformatics* **5**, 154 (2004).
27. M.D. Dyer, T.M. Murali and B.W. Sobral, *Bioinformatics* **23**, i159-66 (2007).
28. F.P. Davis *et al.*, *Protein Sci* **16**, 2585-96 (2007).
29. L. Breiman, *Random Forests*, in *Machine Learning*. 2001. p. 5-32.
30. A. Ben-Hur and W.S. Noble, *BMC Bioinformatics* **7 Suppl 1**, S2 (2006).
31. M. Ashburner *et al.*, *Nat Genet* **25**, 25-9 (2000).
32. D.J. Watts and S.H. Strogatz, *Nature* **393**, 440-2 (1998).
33. L.C. Freeman, *Sociometry* **40**, 35-41 (1977).
34. P. Puntervoll *et al.*, *Nucleic Acids Res* **31**, 3625-30 (2003).
35. M. Matsubara *et al.*, *Protein Sci* **14**, 494-503 (2005).
36. R.R. Baeza-Yates and B.d.A. Neto, *Modern information retrieval*. 1999, New York Harlow, England: ACM Press; Addison-Wesley.
37. S. Bauer *et al.*, *Bioinformatics* **24**, 1650-1 (2008).
38. S. Grossmann *et al.*, *Bioinformatics* **23**, 3024-31 (2007).
39. H.B. Fraser *et al.*, *Science* **296**, 750-2 (2002).
40. H. Jeong *et al.*, *Nature* **411**, 41-2 (2001).
41. H. Liang and W.H. Li, *Trends Genet* **23**, 375-8 (2007).
42. H. Yu *et al.*, *PLoS Comput Biol* **3**, e59 (2007).
43. M.A. Calderwood *et al.*, *Proc Natl Acad Sci U S A* **104**, 7606-11 (2007).
44. M.D. Dyer, T.M. Murali and B.W. Sobral, *PLoS Pathog* **4**, e32 (2008).
45. D.E. Ott, *Rev Med Virol* **18**, 159-75 (2008).
46. S. Mathivanan *et al.*, *Nat Biotechnol* **26**, 164-7 (2008).
47. H. Kato *et al.*, *Genes Dev* **6**, 655-66 (1992).
48. D.J. Mangelsdorf *et al.*, *Cell* **83**, 835-9 (1995).
49. C.H. Wu *et al.*, *Nucleic Acids Res* **34**, D187-91 (2006).
50. G. Nieto *et al.*, *J Steroid Biochem Mol Biol* **89-90**, 199-207 (2004).
51. J. Nevado *et al.*, *J Mol Endocrinol* **38**, 587-601 (2007).