Molecular Systems Biology 5; Article number 303; doi:10.1038/msb.2009.54
Citation: ***Molecular Systems Biology*** **5:303**

www.molecularsystemsbiology.com

molecular systems biology

REPORT

Network-assisted protein identification and data interpretation in shotgun proteomics

Jing Li[1], Lisa J Zimmerman[2], Byung-Hoon Park[3], David L Tabb[1,2], Daniel C Liebler[1,2] and Bing Zhang[1,*]

[1] Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA, [2] Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN, USA and [3] Oak Ridge National Laboratory, Oak Ridge, TN, USA
* Corresponding author. Department of Biomedical Informatics, Vanderbilt University School of Medicine, 2209 Garland Avenue, Nashville, TN 37232-8340, USA. Tel.: +1 161 593 600 90; Fax: +1 161 593 614 27; E-mail: bing.zhang@vanderbilt.edu

Received 18.2.09; accepted 7.7.09

Protein assembly and biological interpretation of the assembled protein lists are critical steps in shotgun proteomics data analysis. Although most biological functions arise from interactions among proteins, current protein assembly pipelines treat proteins as independent entities. Usually, only individual proteins with strong experimental evidence, that is, confident proteins, are reported, whereas many possible proteins of biological interest are eliminated. We have developed a clique-enrichment approach (CEA) to rescue eliminated proteins by incorporating the relationship among proteins as embedded in a protein interaction network. In several data sets tested, CEA increased protein identification by 8–23% with an estimated accuracy of 85%. Rescued proteins were supported by existing literature or transcriptome profiling studies at similar levels as confident proteins and at a significantly higher level than abandoned ones. Applying CEA on a breast cancer data set, rescued proteins coded by well-known breast cancer genes. In addition, CEA generated a network view of the proteins and helped show the modular organization of proteins that may underpin the molecular mechanisms of the disease.
Molecular Systems Biology **5:** 303; published online 18 August 2009; doi:10.1038/msb.2009.54
Subject Categories: proteomics; computational methods
Keywords: clique; data interpretation; protein identification; protein interaction network; shotgun proteomics

Introduction

Shotgun proteomics has emerged as a powerful technology for protein identification in complex samples with remarkable applications in elucidating cellular and subcellular proteomes (Foster *et al*, 2006; Kislinger *et al*, 2006), mapping protein interaction networks (Gavin *et al*, 2006; Krogan *et al*, 2006), and discovering disease biomarkers (Decramer *et al*, 2006; Whiteaker *et al*, 2007). In a typical shotgun proteomics experiment, proteins in a complex mixture are digested by sequence-specific enzymes and the resulting peptides are analyzed by tandem mass spectrometry (MS/MS). Next, MS/MS data acquired from the analyses are processed to identify peptides that gave rise to observed spectra. Finally, proteins are inferred based on peptide identifications and reported.

As proteins are the fundamental units of proteomes, inferring proteins from identified peptides is a critical step in shotgun proteomics (Nesvizhskii and Aebersold, 2005). For each identified peptide, one could add all matched precursor proteins to a maximal protein list. This list comprises the maximal number of possible proteins from the searched database that could explain the observed peptides and may exist in the original sample. In reality, this naïve assembly significantly exaggerates the actual number of proteins in the sample (Nesvizhskii and Aebersold, 2005; Zhang *et al*, 2007).

To ensure the reliability of protein identification, existing protein assembly pipelines usually eliminate a large number of possible but non-confident proteins, including those supported by single peptide and those without distinct peptide evidence (Nesvizhskii and Aebersold, 2005; Zhang *et al*, 2007). However, this conservative assembly may eliminate more than half of all possible proteins, including some truly present proteins that could contribute to the systematic understanding of the biological systems. Indeed, it may introduce a significant

bias against detection of biologically important components of signaling networks, which are often in low abundance and detected by only one peptide. In biomarker studies, conservative assembly may prevent us from identifying important biomarker candidates. Statistical modeling approaches have been proposed to tackle single-hit protein identifications (Nesvizhskii *et al*, 2003; Higdon and Kolker, 2007); however, none of the existing tools is able to handle proteins without distinct peptide identifications. Rescuing true protein identifications from a list of non-confident but possible proteins is still a largely unsolved problem that is extremely challenging, based solely on the observed peptides.

In the current protein assembly pipelines (Nesvizhskii *et al*, 2003; Yang *et al*, 2004; Zhang *et al*, 2007), proteins are considered as independent entities. Nevertheless, accumulating evidence suggests that most biological functions arise from interactions among proteins, and a discrete biological function can only rarely be attributed to an individual protein (Hartwell *et al*, 1999). Recent availability of large-scale protein interaction networks provides an opportunity to investigate shotgun proteomics data at a systems level by taking into consideration the functional relationship among proteins. Previously, protein interaction networks have been used successfully in the prediction of protein functions (Chen and Xu, 2004; Sharan *et al*, 2007), prioritization of candidate disease genes (Oti *et al*, 2006), and classification of cancer metastasis (Chuang *et al*, 2007). In this study, we described a protein interaction network-assisted approach to improve protein identification in shotgun proteomics. Our approach was based on the general concept that proteins involved in the same biological process or pathway tend to lie close to one another in the protein interaction network (Sharan *et al*, 2007).

Using a yeast cell culture data set generated in this study and a published mouse organ data set containing data from brain, placenta, and lung tissues (Kislinger *et al*, 2006), we showed that proteins confidently identified in a specific sample tended to form tightly connected sub-networks in a protein interaction network. These sub-networks might represent key molecular entities that integrate multiple proteins to carry out cellular functions. As cliques in a protein interaction network are sub-networks in which all proteins are pairwisely connected, we attempted to enumerate cliques from protein interaction networks and use the information to improve protein identification. Specifically, we hypothesized that an eliminated but possible protein is more likely to be present in the original sample if it is a member of a sub-network (clique) for which other members have been confidently identified in the same sample. Through simulation studies, we showed that our approach was effective in protein rescue and that this approach outperformed other network-assisted approaches in accuracy and robustness. Application of this approach on the mouse organ data set significantly increased protein identification, and the rescued proteins were well supported by existing literature or transcriptomic studies. Finally, we used a published mouse breast cancer data set (Whiteaker *et al*, 2007) to illustrate that the network-assisted approach not only improved protein identification but also facilitated the biological interpretation and systems level understanding of lengthy protein lists.

Results

Proteins identified in a specific sample form tightly connected sub-networks

Individual shotgun proteomics data sets were processed using the software IDPicker, which combines two-peptide filtering with parsimony analysis in protein assembly (Zhang *et al*, 2007). In the yeast data set, 934 confident proteins were identified. In the mouse organ data set, 1407, 1396, and 1741 confident proteins were identified in the brain, placenta, and lung, respectively.

Proteins identified in the yeast data set were mapped to the yeast protein interaction network (YPIN), which included 5665 proteins and 126 127 interactions. Two mouse protein interaction networks were used to map mouse proteins. The first version (MPIN1) included only literature-supported interactions in mouse and human proteomes collected from public databases, in which human proteins were mapped to mouse orthologs. MPIN1 covered 9776 proteins and 69 470 interactions. As MPIN1 had limited coverage, computationally predicted interactions from Xia *et al* (2006) were appended to MPIN1 to generate the second version of mouse protein interaction network (MPIN2), which covered 12 271 proteins and 236 675 interactions.

A protein interaction network derived from publicly available protein interaction databases includes all known interactions observed in various cellular types and conditions. Therefore, proteins identified in a specific sample under a specific condition will only occupy part of the interaction network. As these proteins are supposed to be functionally related, we hypothesized that they are not randomly distributed on the protein interaction network, instead, they may form tightly connected sub-networks. To test this hypothesis, we constructed sub-networks of proteins identified in each sample to investigate the organization of these proteins using clustering coefficient analysis. The clustering coefficient of a vertex quantifies how well connected the neighborhood of the vertex is. If the neighborhood is fully connected, the clustering coefficient is 1 and a value close to 0 means that there are hardly any connections in the neighborhood. The clustering coefficient of a network is the average clustering coefficient of all vertices in the network. It characterizes the overall tendency of vertices in a network to form clusters or groups (Barabasi and Oltvai, 2004). A high clustering coefficient for a network is an indication of the presence of densely connected neighborhoods in the network (Watts and Strogatz, 1998). As expected, sub-networks of proteins identified in each sample had significantly higher clustering coefficients than the average clustering coefficients obtained from 1000 random sub-networks generated by randomly sampling the same numbers of proteins from corresponding full protein interaction networks (Supplementary Figure 1).

It is worth noting that proteins in the confident protein lists tended to have higher vertex degrees and clustering coefficients in the full protein interaction network. This may have a biological explanation, as essential proteins tend to have higher vertex degrees (Jeong *et al*, 2001) and thus they are likely to be included in the expressed protein lists in all samples. This also may reflect bias in the current shotgun

technology and protein interaction networks, because proteins that are easily identified by the shotgun technology tend to have higher expression levels. Abundant proteins are easier to study and thus show more connections in the current protein interaction networks. Therefore, we further tested whether the higher clustering coefficients of the sub-networks were simply an effect of the topological characteristics of constituent proteins in the full protein interaction networks. Specifically, we divided proteins in the full protein interaction network into nine topological bins based on their degrees and clustering coefficients as described in the Materials and methods section. When we selected random proteins for generating random sub-networks for a sample, for each protein identified in the sample, we chose randomly one protein from the same topological bin. As a result, proteins selected for the random sub-networks had similar degrees and clustering coefficients as the experimentally identified proteins. We refer to these sub-networks as topology-matched random sub-networks. As depicted in Supplementary Figure 1, although these topology-matched random sub-networks tended to have higher clustering coefficients than the completely random sub-networks, they did show significantly lower clustering coefficients than the sub-networks of real proteins. Even in the closest case of the yeast data set, the clustering coefficient of the real sub-network exceeded that of the topology-matched random ones by 8 s.d. (P=9.75e–16).

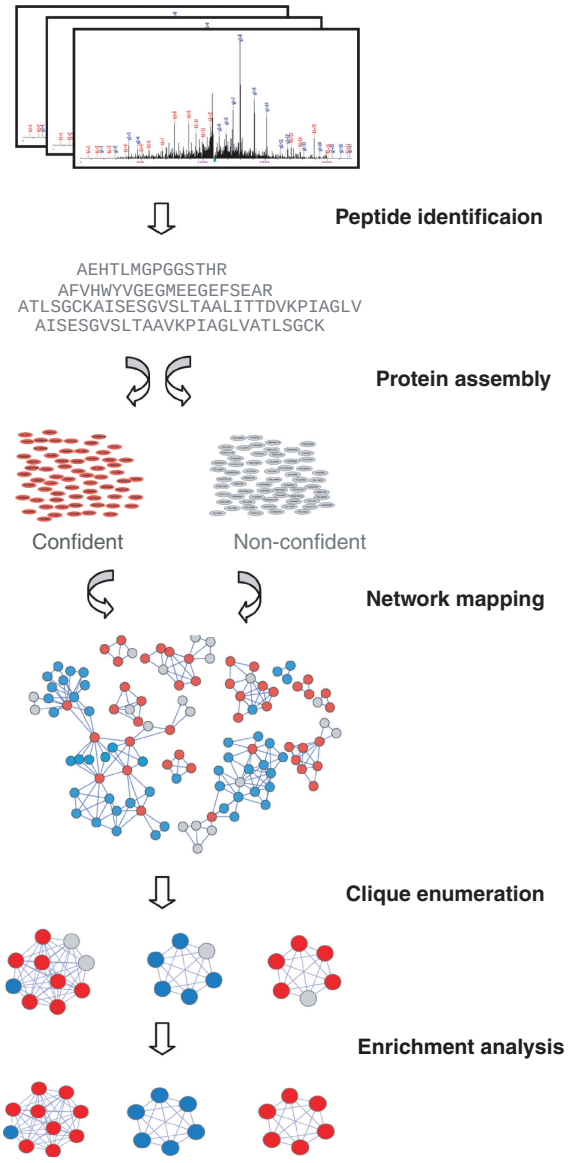
These results suggest that proteins identified in a specific sample are not randomly distributed on the protein interaction network; instead, they tend to form tightly connected sub-networks. As cliques in a protein interaction network are sub-networks in which all proteins are pairwisely connected, we attempted to enumerate cliques from protein interaction networks and use the information to improve protein identification in shotgun proteomics.

Overview of the clique-enrichment approach

To ensure a low protein false discovery rate (FDR), existing protein assembly pipelines eliminate a large number of non-confident but possible proteins. Using the yeast cell culture data set as an example, among the 1742 proteins in the maximal protein list, only 54% were found to be confident proteins according to the rules implemented in IDPicker, whereas 46% were non-confident proteins and thus were eliminated. Non-confident proteins include those supported by single peptide and those without distinct peptide evidence.

To highlight proteins that merit rescue from elimination, we have developed a clique-enrichment approach (CEA), which is illustrated in Box 1. CEA is based on the assumption that an eliminated protein is more likely to be present in the original sample if it is a member of a clique for which other members have been confidently identified in the same sample. First, peptide identification and protein assembly are processed using standard methods; here we employed MyriMatch (Tabb *et al*, 2007) and IDPicker (Zhang *et al*, 2007). Proteins in the maximal protein list are grouped into confident proteins and non-confident proteins after protein assembly, and then mapped to the protein interaction network. In the network, vertices representing confident proteins are labeled as positive (red), vertices representing proteins with no experimental evidence are labeled as negative (blue), and vertices representing non-confident proteins are unlabeled (gray). The next step enumerates all maximal cliques from the protein interaction network. A maximal clique is a clique that is not part of any other larger cliques, that is, inclusion of any other vertex to a maximal clique will violate its completeness. We adopted a graph–theoretic maximal clique finding algorithm (Zhang *et al*, 2008) for this study. For each identified clique, an enrichment score derived from the Fisher's exact test is used to

Box 1 Workflow of the clique-enrichment approach (CEA) for protein identification in shotgun proteomics



Box 1 After peptide identification and protein assembly, proteins in the maximal protein list are grouped into confident proteins (red) and non-confident proteins (gray), and then mapped to the protein interaction network. Proteins absent from the maximal protein list are considered as negative proteins (blue). Cliques are enumerated from the network and evaluated for the enrichment of confident proteins. All non-confident proteins that coexist in a clique enriched with confident proteins are thus rescued and added to the final list, whereas others are discarded.

evaluate the enrichment of confident proteins in the clique. All non-confident proteins that coexist in a clique enriched with confident proteins are thus rescued and added to the final list, whereas others are discarded. The enrichment threshold can be set to achieve desired sensitivity and specificity using cross-validation as described below.

CEA is effective and robust

We used 10-fold cross-validation (see Materials and methods) to evaluate the performance of the proposed method. We first evaluated the performance of CEA in the yeast cell culture data set using YPIN. We defined the gold standard positive set as 834 proteins in YPIN that were confidently identified in the data, and the gold standard negative set as the 4049 proteins that had no experimental evidence according to the data. The 783 non-confident proteins were kept in the network but excluded in the cross-validation test. As shown in the solid red ROC curve in Figure 1A, CEA achieved a specificity (1 false positive rate) of 0.91 or an accuracy (proportion of true results in the predictions) of 0.85 with a sensitivity (true positive rate) of 0.55.

For comparison, we repeated the validation using random networks with the same number of vertices and edges generated by the Erdos–Renyi (ER) model (Erdos and Renyi, 1959). The ROC curve generated from ten ER random networks lay on the diagonal (Supplementary Figure 2), indicating little discrimination power. As the ER random networks do not maintain important topological properties, such as clustering coefficient and degree distribution of the real protein interaction networks, we further repeated the validation with random networks generated through vertex label redistribution, in which the network topological properties were preserved. The ROC curve generated from ten vertex label redistribution random networks also lay on the diagonal (Supplementary Figure 2). These results suggest that the relationship among proteins embedded in the real protein interaction network is critical for achieving good rescue performance.

The performance of CEA was compared with other network-assisted prediction methods. Protein interaction network-based prediction has been an active research topic in other areas such as protein function prediction (Sharan *et al*, 2007) and algorithms developed in those studies could be adopted in our context. Neighbor-voting (NV) is the simplest and most direct method for network-based predictions (Sharan *et al*, 2007), in which the class of an unlabeled vertex is inferred based on the labels of its immediate neighbors. In contrast to NV, which uses only local information, the Hopfield method takes into account the full topology of the network and assigns the classes of the unlabeled vertices so as to minimize the number of edges connecting vertices with different classes (Karaoz *et al*, 2004). We compared the performance of these two methods to that of the CEA through 10-fold cross validation using the yeast cell culture data set. As shown in the solid ROC curves in Figure 1A, the Hopfield method (green) performed very similar to NV (black), whereas CEA (red) clearly outperformed both of them. At the specificity of 90%, the sensitivities for NV, Hopfield, and CEA were 48, 47, and 56%, respectively, whereas the accuracies were 82.8, 82.6, and 84.4%, respectively.

As shotgun proteomics is susceptible to false negatives, we further tested the robustness of the three methods to false negatives by moving 10% of the confident proteins to the negative set. As shown in the dashed ROC curves in Figure 1A, CEA was very robust to the false negatives. In contrast, NV and Hopfield were obviously disturbed by the false negatives.

To test the performance of CEA on mammalian proteomes, in which the network coverage was not as extensive as in yeast, CEA was applied on the mouse organ data set using two versions of protein interaction networks, MPIN1 and MPIN2.

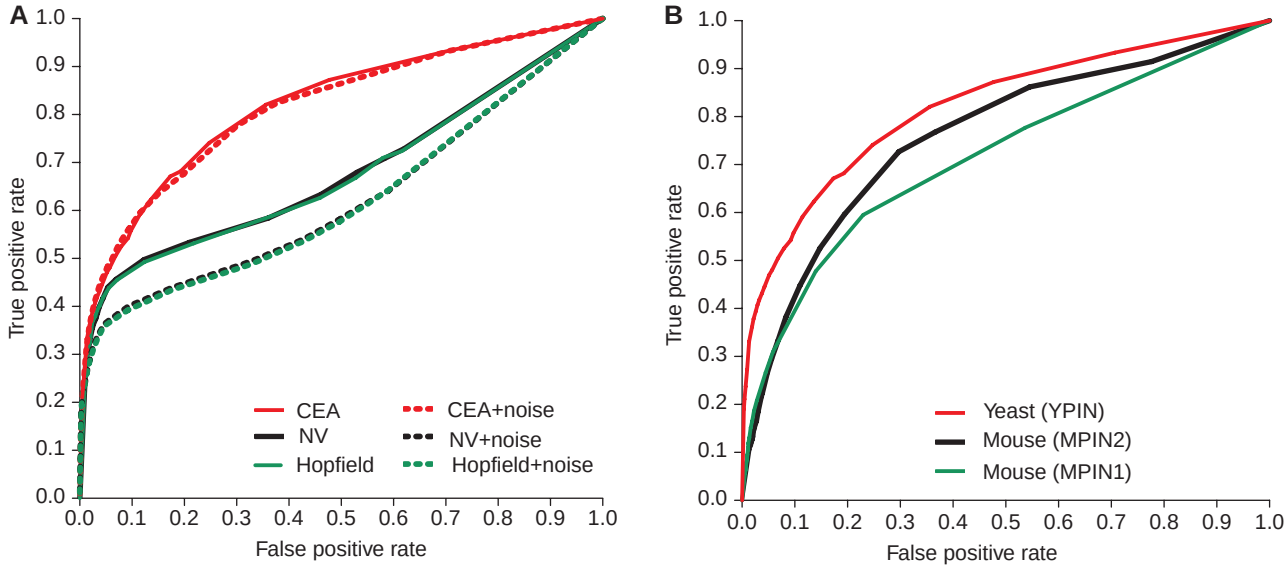


Figure 1 ROC curves from cross-validation studies. (**A**) Comparison among three network-assisted methods: clique-enrichment approach (CEA, red), neighbor-voting (NV, black), and Hopfield (green). Cross-validations using data sets with manually introduced 10% noise are shown in dashed curves. (**B**) ROC curves from the yeast data set (red) and the mouse organ data set (green and black). YPIN: yeast protein interaction network. MPIN1: mouse protein interaction network 1 that includes only literature-supported interactions. MPIN2: mouse protein interaction network 2 that includes both literature-supported and computationally predicted interactions.

Figure 1B compares the performance of CEA in the mouse organ data set using the two versions of networks (black and green) with performance in the yeast data set (red). For the mouse organ data set, the average performance in brain, placenta, and lung samples was plotted. Results for individual tissues are available in Supplementary Figure 3. Clearly, the performance of CEA in the mouse organ data set was not as good as that in the yeast cell culture data set. The area under curve (AUC) based on MPIN2 (black) is better than that based on MPIN1 (green), which suggests the value of improved network coverage. As both networks gave similar sensitivity of around 45% at the specificity of 90%, MPIN1 with literature-supported interactions was used in the following studies.

CEA improves protein identification

After the cross-validation studies described above, CEA was applied to rescue non-confident proteins in the real datasets. The enrichment threshold was specified in each data set separately to achieve an accuracy of 85% based on the cross validation results. For the yeast cell culture data set, 194 of the 783 non-confident proteins in the network were rescued, increasing overall protein identifications by 21% compared with the conservative assembly produced in IDPicker. For the mouse organ data set, 171, 156, and 181 possible proteins were rescued in the brain, placenta, and lung samples, respectively, corresponding to 12, 11, and 10% increases in protein identifications in each organ proteome.

To provide further assessment of the reliability of the rescued proteins, we evaluated the rescued proteins in different organs using relevant data in microarray and EST library studies, as well as through publications indexed in PubMed. Figures 2A–C illustrate the percentage of rescued proteins with different levels of support from the three information resources in the brain, placenta, and lung, respectively. On average, 66% of the rescued proteins were supported by microarray data, 78% were supported by the EST libraries, and 77% were presented in publications on corresponding organs. If we combine different information sources, 49% of the rescued proteins were supported by all of the three information resources, 77% were supported by at least two resources, and 94% were supported by at least one resource.

To evaluate the significance of the support ratios, we further compared the ratios for the rescued proteins with those for all annotated mouse proteins, confident proteins, and un-rescued non-confident proteins. As each information resource used for the evaluation may also have false positives and false negatives, we analyzed the percentage of proteins in each protein set that was supported by, at least, two information resources. As shown in Figure 2D, the support ratios for confident proteins were significantly higher than those for all annotated proteins (*P*-values in Fisher's exact test are 4.90e–90, 1.68e–71, and 4.25e–125 for brain, placenta, and lung, respectively). Interestingly, rescued proteins showed comparable levels of support as confident proteins (*P*-values are 0.51, 0.83, and 0.13 for brain, placenta, and lung, respectively). However, rescued proteins had significantly higher support ratios than un-rescued proteins (*P*-values are 5.28e–7, 2.75e–10, 5.85e–7, for brain, placenta, and lung, respectively). We also carried out the analyses for each information resource separately. The same trend was observed for all comparisons (Supplementary Table 1). These results show that proteins rescued by CEA are reliably identified.

CEA reveals disease-related sub-networks

Finally, we applied CEA on a shotgun proteomics data set comparing tumor and normal mammary tissues from a mouse model of breast cancer. Cross-validation results in this data set were similar to those in the mouse organ data sets (Supplementary Figure 4). CEA increased protein identification by 8 and 23% in the tumor and normal tissues, respectively. Among the 95 rescued non-confident proteins in the tumor tissue, 95 and 33% had been reported in cancer- and breast cancer-related publications. These support levels were significantly higher than those for all annotated proteins (*P*-values in Fisher's exact test are 1.25e–7, 4.20e–4 respectively). Rescued proteins included products from some well-known breast cancer genes, such as *Ctnnb1* and *Top1* (Schroeder *et al*, 2002; Zhao *et al*, 2003; Schlange *et al*, 2007; Yasmeen *et al*, 2007).

As CEA focuses on cliques instead of individual proteins, it provides a logical framework to compare proteomics data sets at the sub-network level. Identified maximal cliques were highly overlapping, which may reflect the involvement of one protein in multiple sub-networks, the dynamic arrangement of the sub-networks, and the incompleteness of protein interaction networks. Overlapping modular structure has been observed in various types of networks, including protein interaction networks, and software has been developed to merge highly overlapping cliques into larger tightly connected sub-networks to show a higher level organization of the networks (Adamcsek *et al*, 2006). To gain a higher level understanding of the difference between tumor and normal tissues, we merged highly overlapping cliques comprising confident and rescued proteins into tightly connected sub-networks using the software, Cfinder (Adamcsek *et al*, 2006). Fourteen cancer-specific sub-networks that comprise only cancer-specific proteins and three normal-specific sub-networks were identified and indicated by different vertex colors in Figure 3 and Supplementary Figure 5.

As shown in Figure 3, 97% of the proteins in the cancer-specific sub-networks had been reported in cancer-related publications (middle vertex size), with 47% in breast cancer-related publications (large vertex size). Interestingly, six of the cancer-specific sub-networks contained four rescued non-confident proteins (triangle vertices). Three of the four rescued proteins had been reported in breast cancer-related publications and all of them had been reported in cancer-related publications. Some proteins in the cancer-specific sub-networks had not been reported in any cancer studies (small vertex size). These proteins may be good candidates for further investigation.

We used Gene Ontology (GO) enrichment analysis (Zhang *et al*, 2005) to identify biological processes associated with the cancer-specific sub-networks. All of the sub-networks showed high functional homogeneity with a Bonferroni-adjusted *P*-value <0.01 in at least one of the biological process categories (Supplementary Table 2). The most enriched GO biological process categories for the sub-networks are labeled

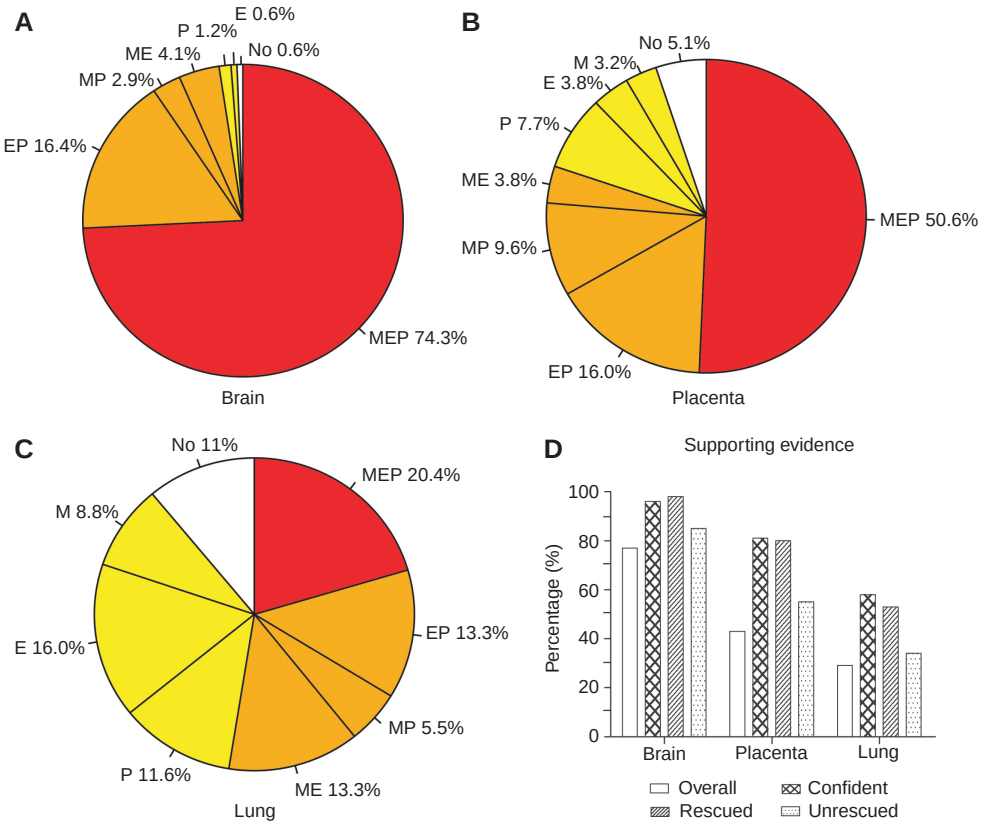


Figure 2 Evaluation of the rescued proteins using relevant gene expression data and publications. (**A**–**C**) For the proteins rescued by the clique-enrichment approach (CEA) in mouse brain, placenta, and lung, relevant data sets in microarray (M), EST library studies (E), and publications in PubMed (P) were investigated for supporting evidence. Red, orange, yellow, and white correspond to support from three, two, one, or zero information resources, respectively. (**D**) Percentage of proteins in all annotated mouse proteins, confident proteins, rescued non-confident proteins, and un-rescued non-confident proteins that are supported by, at least, two information resources in brain, placenta, and lung, respectively.

in Figure 3. These sub-networks corresponded to important biological processes involved in tumor biogenesis and progression, such as 'apoptosis', 'cell adhesion', and 'Wnt receptor signaling pathway' etc. For comparison, we also carried out the GO enrichment analysis for all cancer-specific proteins. As this list was more functionally heterogeneous, we were only able to identify broader categories, such as 'intracellular transport', 'cellular component organization and biogenesis', 'translation' etc (Supplementary Figure 6). Through organizing functionally related proteins together using protein interaction networks, CEA was able to show important but more specific biological processes that involve limited number of proteins.

Discussion

Proteins that are not confidently identified based on multiple peptide identifications and parsimonious assembly in shotgun proteomics are usually eliminated from further consideration. Complete elimination of these possible proteins ensures higher specificity, but sacrifices sensitivity in protein identification. We showed that proteins identified in a specific sample were not randomly distributed on the protein interaction network; instead, they tended to form tightly connected sub-networks. This result suggests that the relationship among proteins embedded in a protein interaction network could provide additional evidence for proteins that are eliminated owing to insufficient experimental evidence. The CEA, proposed in this study, incorporated protein interaction network information and increased protein identification sensitivity while maintaining reasonable accuracy. Indeed, the support levels from independent data sources for rescued proteins were comparable with those for confident proteins (Figure 2D).

Compared with other network-assisted prediction methods, such as NV and Hopfield, CEA proved more effective and robust in our study, which can be explained by its ability to capture the modular architecture of protein interaction networks. Although all three methods are based on the evaluation of neighborhood enrichment, NV and Hopfield do not investigate proteins in a modular context. Instead, all interacting proteins are considered equally and simultaneously. A protein interaction network only represents a collection of possible interactions under many different conditions. Although a protein might be involved in many modules, not all of them are required for a given condition. The evidence of presence for one of them is enough to infer the presence of the protein. Considering the dynamic modular organization of the network, CEA is focused on the most enriched clique and gains sensitivity. However, even under specific conditions, one protein can be involved in multiple

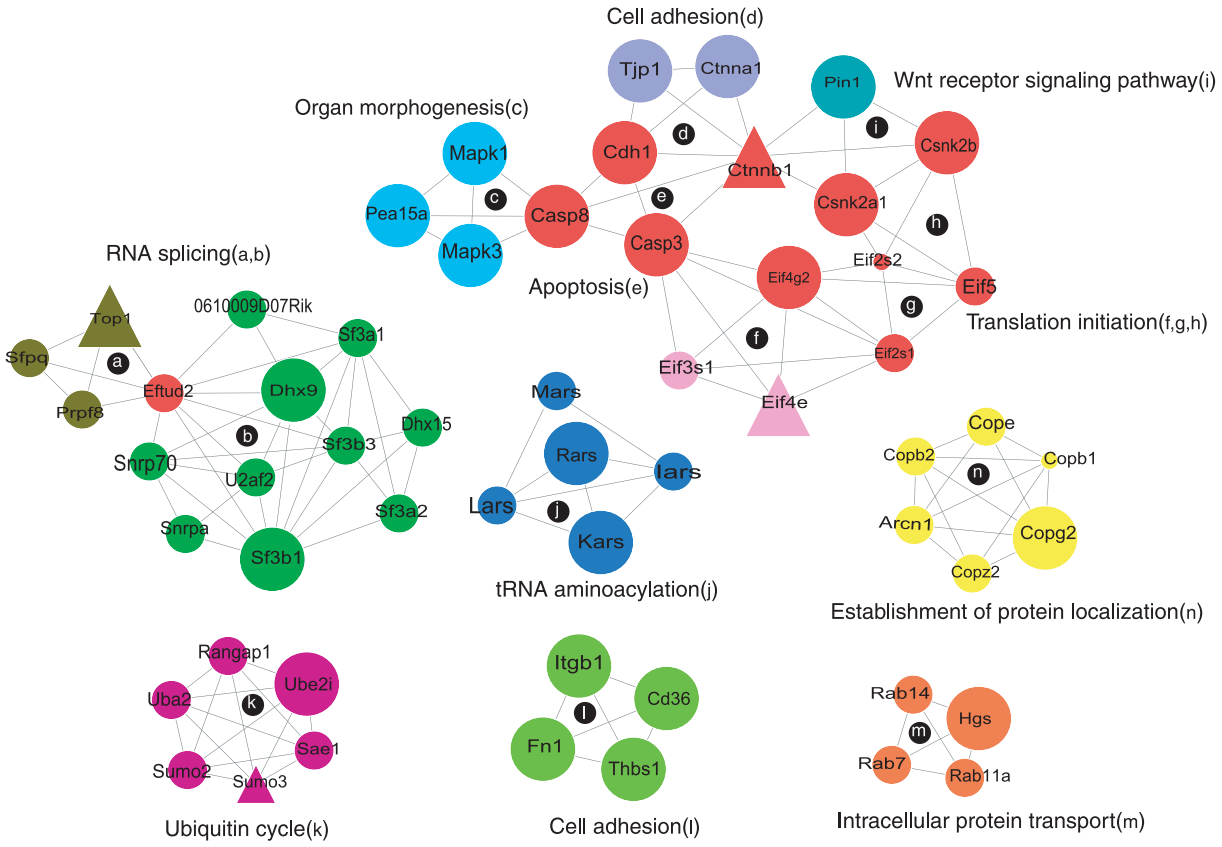


Figure 3 Breast cancer specific sub-networks. Different sub-networks are shown in different colors and identified by IDs from 'a' to 'n'. Proteins shared by multiple sub-networks are colored in red. The most enriched Gene Ontology (GO) biological process annotations for each sub-network are labeled. The IDs accompanying the GO annotations match those of corresponding sub-networks. Triangle vertices represent the proteins rescued by the clique-enrichment approach (CEA). Vertex size represents different levels of publication support: the large size indicates support from breast cancer-related publications, the middle size indicates support from cancer-related publications, and the small size indicates no support from existing cancer-related publications.

modules to carry out different functions. False negative identifications in one module will not necessarily affect other modules. Given the multifunctional nature of proteins, CEA gained robustness by evaluating all possible cliques separately.

In the data sets tested in this study, CEA increased the number of identified proteins by 8–23% without additional proteomic analyses. We want to point out that the improvement was calculated based on the highly confident assembly requiring a peptide FDR of less than 0.05, more than two distinct peptides, and parsimony analysis. The improvement over other protein assembly methods will be different. However, as CEA uses information completely independent of the spectral data, it will complement any method using only spectral data. Despite its promise as a useful tool for protein identification in shotgun proteomics, there are limitations to the performance of CEA based on the following considerations.

First, CEA is dependent on network coverage and quality. As illustrated in Figure 1B, the performance of CEA in the yeast cell culture data set was clearly superior to that in the mouse organ data set. This can be explained by the lower coverage and quality of the mouse protein interaction networks. It is estimated that only ~10% of the human protein interactions are currently known (Hart *et al*, 2006). As the small mouse network was largely derived from the human network, its coverage also was likely to be very low. Although the large mouse network increased the AUC, the major increase was in the region where the specificity fell below 90%. This suggests that the network quality is also very important for achieving both high sensitivity and specificity. Even for yeast, existing databases are estimated to contain only ~50% of all interactions (Hart *et al*, 2006). The growing effort in protein interaction network studies and concomitant improvement of network coverage and quality should improve the performance of CEA.

Second, the 'gold standard' negative set nevertheless contained false negatives. CEA treated all proteins for which no experimental evidence exists as a gold standard negative set. Therefore, false negatives in the shotgun proteomics data would certainly affect the quality of this gold standard. Although we showed that CEA was robust to false negatives, we may have underestimated the performance of CEA in cross-validation. CEA should be able to rescue some proteins that were truly present in the original sample, even if they were treated as gold standard negative in the cross validation, which would lead to an underestimation of performance. Any improvements in the technology sensitivity will reduce false negatives and improve the accuracy of performance evaluation.

Third, cliques are perfectly connected sub-networks and they only represent one type of functional modules. A functional module is composed of multiple molecules that work together in a cell as a distinct unit. Although a module may be a single physical entity, that is, a protein complex, it may also be made from a number of separate physical entities, like a signal transduction pathway. Single physical entities usually constitute tightly connected sub-networks in a protein interaction network, and can be identified by maximal clique enumeration or other well-established algorithms (Snel *et al*, 2002; Bader and Hogue, 2003; Spirin and Mirny, 2003; Newman, 2004; Palla *et al*, 2005). It is more challenging to detect modules consisting of separate physical entities in protein interaction networks. However, a few algorithms have been proposed for this purpose (Steffen *et al*, 2002; Scott *et al*, 2006). Extending our clique enrichment approach to a more general module enrichment approach should help rescue proteins working in functional modules other than cliques and improve the sensitivity of protein identification. As shown in Figure 2D, even un-rescued non-confident proteins were better supported by other information resources than all annotated proteins, which suggests that an improved method should be able to rescue more truly present proteins from the non-confident proteins.

Finally, CEA-based protein rescuing can only be applied on biological samples in which protein interaction networks are expected. For example, we do not anticipate that CEA will work in plasma samples. Moreover, although CEA is extremely useful for shotgun proteomics studies that suffer from high false negative rates, such as studies aiming at the identification of biomarkers, cautions need to be taken if false positive identifications are a major concern. As shown in Figure 2D, although the support levels from independent data sources for rescued proteins were comparable with those for confident proteins, the performance might vary from sample to sample.

Besides accurate protein identification, another challenge in shotgun proteomics is the biological interpretation of the lengthy lists of protein identifications. Although functional class enrichment analysis (Subramanian *et al*, 2005; Zhang *et al*, 2005) could potentially facilitate this process, such analysis is usually limited by existing knowledge on pathways and biological processes. GO and KEGG are the most commonly used databases for defining functional classes. However, out of the 22 762 protein-coding human genes annotated by the Ensembl database (version 48), only 3942 genes (17%) were assigned to KEGG pathways, and only 15 086 (66%) genes were assigned to biological processes in the GO database. Moreover, many genes are annotated at a coarse level in GO, such as 'biological regulation', which is not very helpful in the biological interpretation.

Protein interaction networks provide an alternative way to organize and interpret proteins lists. Mapping proteins in a list to a protein interaction network will help reveal the functional relationships among the identified proteins. More importantly, in comparative analysis, this approach makes it possible to compare proteomics data sets at the network level instead of the individual protein level, which may provide a systems level understanding of the difference between two samples. Recently, protein sub-networks have been proposed as biomarkers for the classification of breast cancer metastasis (Chuang *et al*, 2007). Sub-network biomarkers outperform single gene-based biomarkers in both accuracy and robustness (Chuang *et al*, 2007). In this study, we showed that CEA generated a network view of the identified proteins and helped identify sub-networks that were specific to the cancer phenotype. Subsequent functional profiling of the cancer-specific sub-networks provided insights into molecular mechanisms of cancer (Figure 3). Besides cancer-specific sub-networks, a quantitative score may be calculated to further identify sub-networks that are enriched with cancer-specific proteins, and these sub-networks might also be worth further investigating. We will evaluate different scoring methods for incorporating in the CEA workflow in the future.

Our results also showed that CEA could both rescue biologically important proteins and reveal their biological relevance. For example, the well-known breast cancer protein, Ctnnb1, was supported by single peptide identification in the dataset and would have been eliminated by conservative assembly. CEA not only rescued Ctnnb1 but also assigned it to a sub-network with primary function in the Wnt signaling pathway, which was consistent with the essential role of Ctnnb1 (Fodde and Brabletz, 2007; Malanchi *et al*, 2008). Another interesting rescued protein was Top1. The human ortholog, TOP1, is the only known target of the alkaloid camptothecin, from which the potent anticancer agents, irinotecan and topotecan, are derived (Pommier, 2006). Recently, it has been validated that TOP1 is among the very first proteins to respond to camptothecin in human H1299 lung carcinoma cells (Cohen *et al*, 2008). CEA rescued Top1 and assigned it to a sub-network in which all other proteins were involved in RNA splicing. Although the main function of Top1 is generally considered to be the relaxation of transcription-dependent DNA supercoils, it has also been suggested to have a function in transcript maturation, in particular in the splicing process of mRNAs (Soret *et al*, 2003). Indeed, in a proteomic analyses of Top1 protein complexes, 10 of the 36 proteins identified as Top1 interaction partners are involved in RNA splicing (Czubaty *et al*, 2005).

In conclusion, CEA incorporated protein interaction network information, and greatly improved protein identification and data interpretation in shotgun proteomics. It can be easily integrated into routine shotgun proteomics protein assembly pipelines, such as IDPicker, ProteinProphet, and DBParser (Nesvizhskii *et al*, 2003; Yang *et al*, 2004; Zhang *et al*, 2007). A web-based implementation of CEA is available at the following URL: http://bioinfo.vanderbilt.edu/cea.

Materials and methods

Proteomics data sets

Three data sets were used in this study. The yeast cell culture data set was generated in the Ayers Institute at Vanderbilt. A tryptic digest of a *Saccharomyces cerevisiae* was provided by David Bunk (National Institute of Standards and Technology, Gaithersburg, MD) and analyzed at a concentration of 60 ng/ml using 2 ml injection volumes. The yeast digest was analyzed on an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific) equipped with an Eksigent NanoLC AS1 autosampler and Eksigent NanoLC 1D Plus pump, Nanospray source, and Xcalibur 2.0 SR2 instrument control. Peptides were separated on a packed capillary tip (Polymicro Technologies, 100 mm × 11 cm) with Jupiter C18 resin (5 mm, 300 Å, Phenomenex)

using an in-line solid-phase extraction column (100 mm × 6 cm) packed with the same C18 resin using a frit generated with liquid silicate Kasil 1 (Cortes *et al*, 1987), similar to that previously described (Licklider *et al*, 2002). Mobile phase A consisted of 0.1% formic acid and mobile phase B consisted of 0.1% formic acid in acetonitrile. A 184-min gradient was carried out with a 15-min washing period (100% A for the first 10 min followed by a gradient to 98% A at 15 min) to allow for solid-phase extraction and removal of any residual salts. Following the washing period, the gradient was increased to 40% B by 135 min, followed by an increase to 90% B by 150 min and held for 9 min before returning to the initial conditions. Tandem spectra were acquired using a data-dependent scanning mode in which one full MS scan (m/z 300–2000) was acquired on the Orbitrap at a resolution of 60 000, followed by 8 MS/MS scans collected on the LTQ. The data set may be downloaded from ProteomeCommons.org Tranche, https://proteomecommons.org/tranche/, using the following hash: g6HnZFXo7rRKyLJBFFx98lJT3VoThzIT5Lf4iyCI4TZrV0uKKuKFfvpP-izJcjf9mupIeKeOJmLNmn5ZMy4FaLh8zG58AAAAAAAAIZA==. It is also available at our local website: http://www.mc.vanderbilt.edu/msrc/bioinformatics/data.php.

The mouse organ data set was provided by Kislinger *et al* (2006). The study combined subcellular fractionation with exhaustive MS/MS-based shotgun sequencing to examine the protein content in six organs of the laboratory mouse, *Mus musculus*. We focused on the three organs (brain, placenta, and lung) that contained the most identifiable spectra.

The mouse breast cancer data set was provided by Whiteaker *et al* (2007). The study used LC-MS/MS to examine the protein content in tumor and normal mammary tissues from a conditional HER2/Neu-driven mouse model of breast cancer.

Database search and protein assembly

Using Myrimatch version 1.2.9 (Tabb *et al*, 2007), MS/MS spectra were identified against the *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org/) for yeast, and Swiss-Prot (release 53.1) for mouse. In both cases, the reversed version of each protein sequence was appended. Only tryptic peptides were considered. All cysteines were assumed to be carboxamidomethylated, and methionines were allowed to be oxidized. A precursor error of up to 1.25 m/z was permitted, whereas fragment ions were required to fall within 0.5 m/z of their expected locations. Ambiguous identifications that mapped spectra to multiple peptide sequences at equal scores were excluded. Peptide identification and protein assembly was carried out using IDPicker (Zhang *et al*, 2007). Instead of relying on peptide identification score thresholds, IDPicker estimates FDRs from reversed-sequence database search to control the quality of peptide identification. The peptide FDR cutoff was set to 0.05 for all data sets in this study. The two distinct-peptide requirement and parsimony analysis were applied in protein assembly.

Protein interaction networks

Yeast protein interaction network

Protein interaction data were downloaded from BioGRID (http://www.thebiogrid.org/, version 2.0.35 release) and MIPS (Mewes *et al*, 2004) (http://mips.gsf.de/) on 12/20/2007 and integrated. The BioGRID network included 5049 proteins and 69 228 interactions, and the MIPS network contained 5003 proteins and 76 856 interactions. The integrated network was comprised of 5665 proteins and 126 127 interactions.

Mouse protein interaction network

We downloaded literature-supported human and mouse protein interactions from seven databases: HPRD (v7; including HPRD_COMPLEX), DIP, MINT, MIPS, REACTOME, and INTACT (latest updated on 1/10/2008). Genes in human protein interactions were mapped to mouse orthologs according to the mouse–human orthology map from MGI (http://www.informatics.jax.org/orthology.shtml). A literature-supported network (MPIN1) with 69 470 interactions and 9776 proteins were obtained after removing redundancy. In addition, this network was appended with computationally predicted mouse protein–protein interactions (Xia *et al*, 2006) to form a high-coverage network with 12 271 proteins and 236 675 interactions (MPIN2).

Clustering coefficient analysis and network randomization

Clustering coefficient was calculated according to Watts and Strogatz (1998). Random sub-networks were generated by randomly sampling a desired number of vertices from a full network, and any edges that connected two sampled vertices were kept in the random sub-network. To generate topology-matched random sub-networks for a real sub-network, we first assigned all proteins in a full protein interaction network to three bins of equal size based on their degrees, and similarly three bins based on their clustering coefficients. Next, proteins were divided into nine topological bins based on the combination of their degree bin and clustering coefficient bin assignments. Finally, for each protein included in a real sub-network, we chose randomly one protein from the same topological bin to create a topology-matched random sub-network. ER random networks were generated according to the ER model (Erdos and Renyi, 1959). Vertex label redistribution random networks were generated through randomly reshuffling the vertex labels while maintaining the network topology.

Network-assisted approaches

A protein interaction network is represented as an undirected graph $G(V,E)$ that consists of a set of vertices V and a set of edges E. Each vertex represents a protein, whereas each edge represents an interaction between two proteins. We considered three network-assisted approaches to predict the class label of the non-confident proteins.

CEA is a clique-based approach that first identifies all maximal cliques in the network (Zhang *et al*, 2008) and then assigns a label of presence or absence to each clique and associated non-confident proteins based on a statistical scoring algorithm. The enrichment score for a clique c is computed as the following:

$$S_c = -\log_{10}^{f(m,n,j,k)}$$

$$f(m,n,j,k) = \sum_{i=k}^{\min(n,j)} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

where m, n, j, k denote the numbers of all the proteins involved in, at least, one clique; confident proteins involved in, at least, one clique; all proteins in clique c; and confident proteins in clique c, respectively. Clique c is assigned a present label if its enrichment score is higher than a predefined threshold. A non-confident protein is assigned a present label if it is involved in, at least, one clique with a present label.

The NV algorithm is adopted from Karaoz *et al* (2004) as described in the following formula:

$$S_i = \sum_{1 \le j \le n_i} W_{ij} S_j.$$

Here, n_i denotes the number of neighbors of protein i, and s_j denotes the label of the jth neighbor of protein i. For protein j, s_j=1 if it has a positive (present) label, s_j=−1 if it has a negative (absent) label, and s_j=0 if it is unlabelled. w_{ij} is the weight of the edge connecting protein i and protein j, which is set to 1 in this study. A possible protein 'i' is assigned a present label if s_j is higher than a predefined threshold.

The Hopfield algorithm is as described by Karaoz *et al* (2004). Briefly, each positive vertex is assigned a state S_v that equals $+1$, whereas each negative vertex is assigned a state that equals -1. Next, an assignment of -1 and $+1$ states to the unlabeled vertices is sought so as to maximize $\sum_{(i,j)\in E} S_i S_j$. This is achieved by an iterative procedure in which for every unlabeled vertex, in turn, the state of the vertex is changed according to the majority of the states of its neighbors, until satisfactory convergence is reached.

Cross-validation tests

Tenfold cross-validation was used to evaluate the performance of the network-assisted approaches. A gold standard positive set included all confident proteins. For the proteins in an indiscernible protein group (Zhang *et al*, 2007), the protein with the largest number of positive direct interaction partners was also added to the positive set. A gold standard negative set comprised all proteins with no peptide identification. Non-confident proteins were kept in the network but excluded in the cross-validation tests. The data were partitioned into ten equal partitions and each in turn was used for testing, whereas the remainder was used for training. As we had many more negative instances than positive instances in the dataset, stratified 10-fold cross-validation was used to ensure that each fold had roughly the same proportion of class labels as in the original data set. After the cross-validation tests, all proteins in the 10 test sets were ranked together based on their predicted scores. True labels of the proteins were retrieved for ROC curve generation.

Evaluation of the CEA predictions

Existing microarray and EST profiling data sets, and PubMed records were used to evaluate the predictions in the mouse data sets. The microarray gene expression profiling data set (Su *et al*, 2004) was downloaded from the GEO database (Barrett *et al*, 2007) of NCBI (http://www.ncbi.nlm.nih.gov/geo/, GDS592). Expression data on mouse normal lung, placenta, and brain samples were used for the evaluation. Genes were considered to be expressed if they had a present call. EST-based gene expression profiles of mouse normal lung, placenta, and brain were download form CGAP (http://cgap.nci.nih.gov/Tissues). For the PubMed-based evaluation, we first associated publications with various keywords, including 'lung', 'placenta', 'brain', 'cancer', and 'breast cancer', through PubMed search. Next, we downloaded the gene to publication association table Gene2-pubmed from NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA). A Perl script was written to process these two association tables to generate gene lists related to each keyword. Fisher's exact test was used to compare the difference of support between two selected protein lists.

Clique merging, sub-network visualization, and functional evaluation

Maximal cliques were merged to form tightly connected sub-networks using the software Cfinder (Adamcsek *et al*, 2006) and visualized using Cytoscape (Shannon *et al*, 2003). From a list of maximal cliques, CFinder identifies the tightly connected sub-networks by carrying out a standard component analysis of the clique–clique overlap matrix (Palla *et al*, 2005). GO enrichment of the sub-networks was analyzed using WebGestalt (Zhang *et al*, 2005).

Software implementation

CEA was implemented in Perl and PHP. It is available at http://bioinfo.vanderbilt.edu/cea.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Dr Whiteaker and Dr Kislinger for making the mouse organ data set and the mouse breast cancer data set available, respectively. This work was conducted, in part, using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. This work was supported by the National Institutes of Health (NIH)/ National Cancer Institute (NCI) through grant R01 CA126218 and the NCI Clinical Proteomic Technologies Assessment for Cancer program through grant 1U24CA126479.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Adamcsek B, Palla G, Farkas IJ, Derenyi I, Vicsek T (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* **22:** 1021–1023

Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4:** 2

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5:** 101–113

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res* **35:** D760–D765

Chen Y, Xu D (2004) Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res* **32:** 6414–6424

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3:** 140

Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, Sigal A, Milo R, Cohen-Saidon C, Liron Y, Kam Z, Cohen L, Danon T, Perzov N, Alon U (2008) Dynamic proteomics of individual cancer cells in response to a drug. *Science* **322:** 1511–1516

Cortes A, Pfeiffer HJ, Richter BE, Stevens T (1987) Porous ceramic bed supports for fused silica packed capillary columns used in liquid chromatography. *HRC CC J High Resolut Chromatogr* **10:** 446–448

Czubaty A, Girstun A, Kowalska-Loth B, Trzcinska AM, Purta E, Winczura A, Grajkowski W, Staron K (2005) Proteomic analysis of complexes formed by human topoisomerase I. *Biochim Biophys Acta* **1749:** 133–141

Decramer S, Wittke S, Mischak H, Zurbig P, Walden M, Bouissou F, Bascands JL, Schanstra JP (2006) Predicting the clinical outcome of congenital unilateral ureteropelvic junction obstruction in newborn by urinary proteome analysis. *Nat Med* **12:** 398–400

Erdos P, Renyi A (1959) On random graphs. *Publicationes Mathematicae* **6:** 290–297

Fodde R, Brabletz T (2007) Wnt/beta-catenin signaling in cancer stemness and malignant behavior. *Curr Opin Cell Biol* **19:** 150–158

Foster LJ, de Hoog CL, Zhang Y, Zhang Y, Xie X, Mootha VK, Mann M (2006) A mammalian organelle map by protein correlation profiling. *Cell* **125:** 187–199

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M *et al* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440:** 631–636

Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol* **7:** 120

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* **402:** C47–C52

Higdon R, Kolker E (2007) A predictive model for identifying proteins by a single peptide match. *Bioinformatics* **23:** 277–280

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411:** 41–42

Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci USA* **101:** 2888–2893

Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, Rossant J, Hughes TR, Frey B, Emili A (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125:** 173–186

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B *et al* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440:** 637–643

Licklider LJ, Thoreen CC, Peng J, Gygi SP (2002) Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column. *Anal Chem* **74:** 3076–3083

Malanchi I, Peinado H, Kassen D, Hussenet T, Metzger D, Chambon P, Huber M, Hohl D, Cano A, Birchmeier W, Huelsken J (2008) Cutaneous cancer stem cell maintenance is dependent on beta-catenin signalling. *Nature* **452:** 650–653

Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, Warfsmann J, Ruepp A (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32:** D41–D44

Nesvizhskii AI, Aebersold R (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4:** 1419–1440

Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75:** 4646–4658

Newman ME (2004) Fast algorithm for detecting community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **69:** 066133

Oti M, Snel B, Huynen MA, Brunner HG (2006) Predicting disease genes using protein–protein interactions. *J Med Genet* **43:** 691–698

Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435:** 814–818

Pommier Y (2006) Topoisomerase I inhibitors: camptothecins and beyond. *Nat Rev Cancer* **6:** 789–802

Schlange T, Matsuda Y, Lienhard S, Huber A, Hynes NE (2007) Autocrine WNT signaling contributes to breast cancer cell proliferation via the canonical WNT pathway and EGFR transactivation. *Breast Cancer Res* **9:** R63

Schroeder JA, Adriance MC, McConnell EJ, Thompson MC, Pockaj B, Gendler SJ (2002) ErbB-beta-catenin complexes are associated with human infiltrating ductal breast and murine mammary tumor virus (MMTV)-Wnt-1 and MMTV-c-Neu transgenic carcinomas. *J Biol Chem* **277:** 22692–22698

Scott J, Ideker T, Karp RM, Sharan R (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol* **13:** 133–144

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. *Mol Syst Biol* **3:** 88

Snel B, Bork P, Huynen MA (2002) The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci USA* **99:** 5890–5895

Soret J, Gabut M, Dupon C, Kohlhagen G, Stevenin J, Pommier Y, Tazi J (2003) Altered serine/arginine-rich protein phosphorylation and exonic enhancer-dependent splicing in Mammalian cells lacking topoisomerase I. *Cancer Res* **63:** 8203–8211

Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* **100:** 12123–12128

Steffen M, Petti A, Aach J, D'Haeseleer P, Church G (2002) Automated modelling of signal transduction networks. *BMC Bioinformatics* **3:** 34

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* **101:** 6062–6067

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102:** 15545–15550

Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **6:** 654–661

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* **393:** 440–442

Whiteaker JR, Zhang H, Zhao L, Wang P, Kelly-Spratt KS, Ivey RG, Piening BD, Feng LC, Kasarda E, Gurley KE, Eng JK, Chodosh LA, Kemp CJ, McIntosh MW, Paulovich AG (2007) Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. *J Proteome Res* **6:** 3962–3975

Xia K, Dong D, Han JD (2006) IntNetDB v1.0: an integrated protein–protein interaction network database generated by a probabilistic model. *BMC Bioinformatics* **7:** 508

Yang X, Dondeti V, Dezube R, Maynard DM, Geer LY, Epstein J, Chen X, Markey SP, Kowalak JA (2004) DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res* **3:** 1002–1008

Yasmeen A, Bismar TA, Dekhil H, Ricciardi R, Kassab A, Gambacorti-Passerini C, Al Moustafa AE (2007) ErbB-2 receptor cooperates with E6/E7 oncoproteins of HPV type 16 in breast tumorigenesis. *Cell Cycle* **6:** 2939–2943

Zhang B, Chambers MC, Tabb DL (2007) Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res* **6:** 3549–3557

Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* **33:** W741–W748

Zhang B, Park BH, Karpinets T, Samatova NF (2008) From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics* **24:** 979–986

Zhao C, Yasui K, Lee CJ, Kurioka H, Hosokawa Y, Oka T, Inazawa J (2003) Elevated expression levels of NCOA3, TOP1, and TFAP2C in breast tumors as predictors of poor prognosis. *Cancer* **98:** 18–23