Integrating Protein Family Sequence Similarities with Gene Expression to Find Signature Gene Networks in Breast Cancer Metastasis

Sepideh Babaei[1,2], Erik van den Akker[1,3], Jeroen de Ridder[1,2,⋆], and Marcel Reinders[1,2,⋆]

[1] Delft Bioinformatics Lab, Delft University of Technology, The Netherlands
[2] Netherlands Bioinformatics Centre, The Netherlands
[3] Molecular Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands
{J.deRidder,M.J.T.Reinders}@tudelft.nl

Abstract. Finding robust marker genes is one of the key challenges in breast cancer research. Significant signatures identified in independent datasets often show little to no overlap, possibly due to small sample size, noise in gene expression measurements, and heterogeneity across patients. To find more robust markers, several studies analyzed the gene expression data by grouping functionally related genes using pathways or protein interaction data. Here we pursue a protein similarity measure based on Pfam protein family information to aid the identification of robust subnetworks for prediction of metastasis. The proposed protein-to-protein similarities are derived from a protein-to-family network using family HMM profiles. The gene expression data is overlaid with the obtained protein-protein sequence similarity network on six breast cancer datasets. The results indicate that the captured protein similarities represent interesting predictive capacity that aids interpretation of the resulting signatures and improves robustness.

Keywords: protein-to-family distance matrix, protein-to-protein sequence similarity, concordant signature, breast cancer markers.

1 Introduction

Delineating gene signatures that predict cancer patient prognosis and survival is an important and challenging question in cancer research. Over the last few years, the amount of data from breast cancer patients has increased [1] and various methods for inferring prognostic gene sets from these data have been proposed [2], [3]. A major problem in this, which has been identified in several studies already, is that the prognostic signatures have relatively low concordance between different studies [4]. This is apparent from the fact that prediction performance decreases dramatically when prognostic signatures obtained from one dataset are applied to another one [5]. This reveals that a lack of a unified mechanism through which clinical outcome can be explained from gene expression profiles is still a major hurdle in clinical cancer biology.

⋆ Corresponding author.

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 247–259, 2011.

Several studies connect the lack of overlap in the gene expression to insufficient patient sample size [6], the inherent measurement noise in microarray experiments or heterogeneity in samples [5], [7]. A possible remedy is to pool breast cancer datasets in order to capture the information of as many samples as possible in the predictor [4], [8]. More recent efforts address this problem by exploiting knowledge on relations between genes and infer signatures not as individual genes but related groups of genes [9]. Of particular interest is the pioneering work of Chuang *et al.* [10], in which a greedy algorithm is used to detect discriminative subnetworks from protein-protein interaction (PPI) networks.

In order to overcome the drawbacks of a greedy routine to select candidate cancer networks, van den Akker *et al.* [11] proposed a non-greedy method that overlays the PPI network with gene-expression correlation to more accurately determine concordant breast cancer signatures across six independent expression datasets. These studies demonstrate that using additional information results in signatures with a more robust performance across dataset. Additionally, these signatures also turn out to have meaningful biological interpretations, thus providing interesting clues for the underlying molecular mechanisms of metastasis.

In this study, we propose a novel method to derive protein networks based on their functional similarities and use this to extend the van den Akker *et al.* method in order to further improve signature concordance and biological interpretability of breast cancer classification. To this end, we exploit protein sequence similarity since proteins with significant similar sequence are likely to have similar function [12].

Protein sequence similarity is determined using profile hidden Markov models (profile HMM) of protein families released in Pfam database [13]. More specifically, using the HMMER sequence alignment tool [14], we measure a distance (in term of an E-value) between a protein sequence and the HHM profiles constructed from the known families (Fig. 1). Proteins with common ancestors are in close proximity in the family space, sharing one or more protein domains since they are close to the same family profiles. Dimensionality reduction methods are applied to transform the high-dimensional protein-to-family matrix into a more meaningful low-dimensional representation in which only relevant dimensions are retained. The protein-to-protein similarity matrix (PPS) is derived by taking the Euclidean distance between pairs of proteins in the family space. We hypothesize that the captured protein similarity networks express predictive power that can be exploited in a cancer classification task. We evaluated these networks to identify predictive subnetworks in breast cancer metastasis.

2 Material and Methods

2.1 Data Description

In this study, six microarray datasets of breast cancer samples (Table 1) are utilized. The samples are measured on the HG U133 platform, and normalized following van den Akker *et al.* [11]. Briefly, all microarray data is normalized, log2 transformed and summarized per probe set. Duplicated samples are removed.

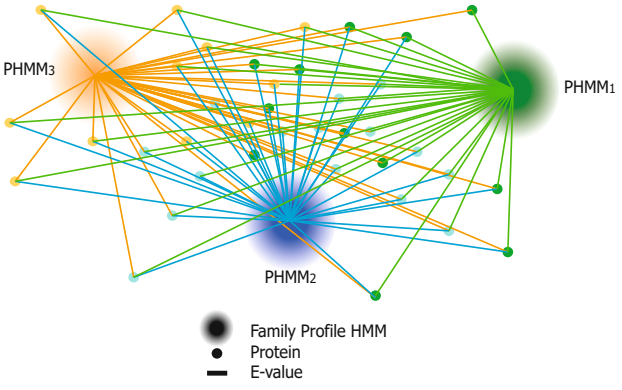

Fig. 1. The PPS is constructed based on profile HMM of protein families. These profiles are applied to scan each of the protein queries to obtain their sequence similarity in terms of an E-value.

The probe sets are mapped to protein-protein interaction networks (PPI) obtained from STRING [15]. Probe sets that do not map to known transcripts or proteins are discarded, resulting in a total of $N = 9236$ probe sets for 1107 samples. The samples are classified into a Poor or Good class according to the prognosis labels (distant metastasis and free survival events, respectively) [11].

Table 1. Summary of collected datasets

Dataset	Accession Code	Poor Sample	Good Sample
Desmedt	GSE7390	31	119
Miller	GSE3494	37	158
Loi	GSE6532	32	107
Pawitan	GSE1456	35	115
Wang	GSE2034	95	180
Schmidt	GSE11121	27	153

2.2 Protein-Protein Sequence Similarity

The protein-to-protein similarities are based on Pfam gene family information. These similarities are calculated as follows (summarized in Fig. 2):

Step 1. Profile hidden Markov models (HMM) for the families in Pfam24 dataset (2009 release, containing 11912 protein families) are constructed using the HMMER sequence alignment tool [14]. Families with less than 10 or more than 100 protein members are removed ($M = 6612$).

Step 2. These profile HMMs are used to scan N protein queries to obtain their sequence similarity to the $M = 6612$ families in terms of an E-value. The $N = 9236$ protein sequences associated with probe sets are taken from the human genes in the Ensemble database [16] (Fig. 1).

Step 3. We next estimate the intrinsic dimensionality of the protein-to-family distance matrix to capture the meaningful structure. Various dimension reduction methods are examined. In particular, classical multi dimensional scaling (MDS) and t-distributed stochastic neighbor embedding (t-SNE) methods are employed to capture a low-dimensional manifold that embeds the proteins in a low family dimensional space [17], i.e. the $M \times N$ matrix containing the distances between the N protein sequences and M protein families is mapped to a $k \times N$ ($k \leqslant M$) dimensional space.

Step 4. The $N \times N$ protein-to-protein similarity matrix (PPS) is extracted from the mapped protein-to-family distance matrix by taking the Euclidean distance between pairs of proteins in the mapped space. Depending on the employed dimension reduction approach, the PPS is referred to as $\boldsymbol{PPS}_{SNE}$ or $\boldsymbol{PPS}_{MDS}$. $\boldsymbol{PPS}_{ORG}$ refres to the protein distance matrix extracted from the non-mapped $M \times N$ matrix as well (i.e. without dimension reduction step).

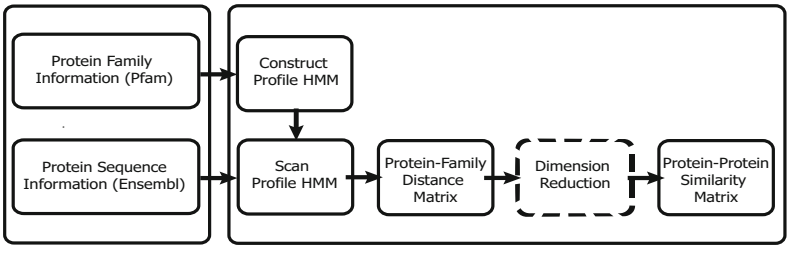

Fig. 2. Flow diagram of capturing the protein-to-protein similarity matrix (***PPS***) framework. Three types of ***PPS*** are extracted: $\boldsymbol{PPS}_{SNE}$, $\boldsymbol{PPS}_{MDS}$ and $\boldsymbol{PPS}_{ORG}$. For the latter the dimension reduction step is skipped.

2.3 Subnetworks Construction

To construct subnetworks we mostly follow van den Akker *et al.* [11]. Briefly, we utilize different types of evidence to create the initial subnetworks including expression correlations, physical protein interactions or protein functional similarities. The $N \times N$ matrix indicating relations between genes is computed by:

$$\boldsymbol{S} = \boldsymbol{C}_B \circ \boldsymbol{G} \circ \boldsymbol{P}_B$$

where "$\circ$" indicates Hadamard product and $\boldsymbol{C}_B$ is an $N \times N$ binary matrix in which each element in the correlation matrix $\boldsymbol{C}$ is set to zero in case it does not exceed threshold (T_{COR}). Grouping matrix $\boldsymbol{G}$ ($N \times N$) refers to the expression clustering matrix computed by hierarchically clustering genes using average linkage and a cluster cut-off value equal to $1 - T_{COR}$. Nonzero values in this matrix indicate the co-membership of a gene pair in a cluster. The binary protein association matrix $\boldsymbol{P}_B$ ($N \times N$) captures the functional relationship between proteins. This matrix is constructed based on the ***PPS*** matrix or the ***PPI*** matrix. In case of the latter, the STRING based interaction confidences, ranging

from 1 to 999, are thresholded using threshold T_{PPI} such that a nonzero value indicates an interaction between the corresponding protein pair. Alternatively, $\boldsymbol{P}_B$ can be constructed from the $\boldsymbol{PPS}$ matrix, by thresholding it with threshold T_{PPS}. A nonzero value in $\boldsymbol{P}_B$ indicates the high similarity of a protein pair.

To assess the association between the gene expression data and breast cancer outcome per captured subnetworks the global test [18] is applied as a summary statistic. Only subnetworks with a global p-value less than T_S are considered as the significant networks. Thereafter, genes within union of significant subnetworks are included in the classifier and used to find out similarity in gene selection between different datasets.

2.4 Signature Concordance

In order to quantify the degree of concordance between signatures derived from different datasets two statistics, the Jaccard index [19] and odds ratio [20], are used. We specified four combinations of attributes for given subnetworks S_i and S_j as the total number of genes that are in i) both S_i and S_j (n_{11}), ii) neither S_i and S_j (n_{00}), iii) only S_i (n_{01}) and iv) only S_j (n_{10}). The Jaccard coefficient measures the degree of overlap between two networks by computing the ratio of the shared attributes between S_i and S_j: $J = \frac{n_{11}}{n_{11}+n_{01}+n_{10}}$.

The odds ratio describes the strength of association between two subnetworks. It is the ratio of the odds of an event occurring in S_i to the odds of it occurring in S_j: $OR = \frac{n_{11}n_{00}}{n_{01}n_{10}}$. Therefore, a high value of both criteria across two different datasets indicates the consistency of selected genes between them.

2.5 Classification Procedure

The predictive performances of the significant subnetworks are examined by training classifiers. More specifically, a nearest mean classifier using cosine correlation as the distance metric is trained on expression values associated with genes in the significant subnetworks. The classifiers are evaluated using the area under ROC curve, AUC metric, to capture the performance over the range of sensitivity and specificity. Two different strategies for training cross-dataset classifiers are evaluated: geneset passing and classifier passing. In geneset passing, the gene selection is based on the significant genes in dataset A while the classifier is trained and tested on the expression value associated with these selected genes in dataset B using 5-fold cross validation. The procedure is repeated 100 times and thus, the reported performance is the average of the AUC among all repeats. The classifier passing routine is carried out based on selecting genes and training the classifier on dataset A while testing on dataset B.

Finally, the datasets are also integrated to assess the classifier performance using five datasets for training while testing on the sixth. In an early integration strategy, five dataset are concatenated and then the statistically significant subnetworks are identified using the global test. Alternatively, for late integration, the gene sets are determined by intersecting the significant subnetworks per dataset.

3 Results and Discussion

We evaluate the derived protein similarity matrix to find out whether it represents biologically meaningful relationships among the proteins by means of a comparison with the Gene Ontology (GO) categorization. The protein-to-family matrix ($k \times N$) is mapped to low-dimensional space using t-SNE method in which $k = 3$ and N is set to ten thousand randomly selected of the all human proteins in Ensembl database [16]. As shown in Fig. 3, the proteins with a common GO term are in close proximity in the PPS space. This shows our measure is indeed capable of capturing function relatedness of proteins and gives confidence that it can aid the construction of concordant and biologically interpretable signatures.

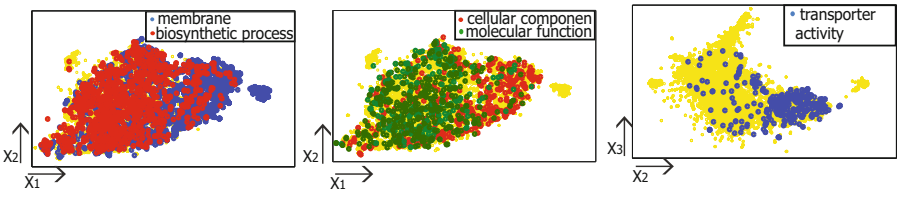

Fig. 3. Visualization of proteins colored based on their GO categorization in the mapped $\boldsymbol{PPS}_{SNE}$ space ($k = 3$). Proteins with a common GO term are closer in the mapped space. Yellow dots indicate all proteins in family space.

Following the outlined procedure, the subnetwork matrix $\boldsymbol{S}$ is created based on expression correlations ($\boldsymbol{C}$) in the six aforementioned datasets independently. by setting $T_{COR} = 0.6$ (following van den Akker *et al.* [11]). We evaluate four different subnetwork matrices $\boldsymbol{S}$ by varying the way in which the protein association term ($\boldsymbol{P}_B$) is calculated. More specifically, we used: $\boldsymbol{PPS}_{SNE}$, $\boldsymbol{PPS}_{MDS}$ and $\boldsymbol{PPS}_{ORG}$ as well as $\boldsymbol{PPI}$ directly. For the latter, the threshold T_{PPI} is set to 500. In case of using the protein similarity measures, T_{PPS} is set to 0.1 quantile of all the values in the corresponding $\boldsymbol{PPS}$ matrix. The global test is

Table 2. Number of significant genes and subnetworks obtained by four different protein association terms ($\boldsymbol{P}_B$) on the six breast cancer datasets. For each method the first column indicates number of significant genes and the percentage among all genes on the array, the second column indicates the number of selected subnetworks and their average size.

Dataset	$\boldsymbol{PPI}$		$\boldsymbol{PPS}_{ORG}$		$\boldsymbol{PPS}_{MDS}$		$\boldsymbol{PPS}_{SNE}$	
	#Genes(%)	#Net(mean)	#Genes(%)	#Net(mean)	#Genes(%)	#Net(mean)	#Genes(%)	#Net(mean)
Desmedt	117(1.2)	37(5.3)	44(0.4)	18(4.4)	84(0.9)	30(4.9)	85(0.9)	34(4.7)
Miller	306(3.3)	56(7.5)	130(1.4)	35(5.8)	207(2.2)	61(5.5)	192(2.1)	54(5.8)
Loi	819(8.8)	113(9.3)	556(6)	119(6.9)	776(8.4)	157(7.2)	751(8.1)	168(6.9)
Pawitan	246(2.6)	52(6.8)	109(1.2)	38(5.2)	206(2.2)	56(5.6)	190(2.1)	56(5.7)
Wang	293(3.1)	72(6.3)	122(1.3)	58(4.4)	226(2.4)	71(5.3)	227(2.4)	70(5.3)
Schmidt	209(2.2)	50(6.2)	99(1.1)	32(5.1)	186(2)	48(6)	190(2.1)	47(6.2)

performed on the obtained matrices $\boldsymbol{S}$ using $T_S = 0.05$ to identify the significant subnetworks. As a result, four different subnetwork matrices for six expression datasets are obtained. The summary of significant genes and subnetworks per dataset are reported in Table 2.

3.1 Cross Study Prediction Evaluation

We examine the classification performances of the discriminative subnetworks to compare the prediction capability of consensus genes selected by various protein association procedures. The expression values of the selected genes are employed to train and test the classifier. We evaluated two classification protocols, geneset passing and classifier passing, on the data resulting from early as well as late integration. The AUC values are given in Fig. 4. From these results we learn that in terms of classifier performance the genesets obtained with both the $\boldsymbol{PPS}$ matrices and the $\boldsymbol{PPI}$ matrix are in the same range, with a perhaps slightly better performance for the $\boldsymbol{PPS}_{ORG}$.

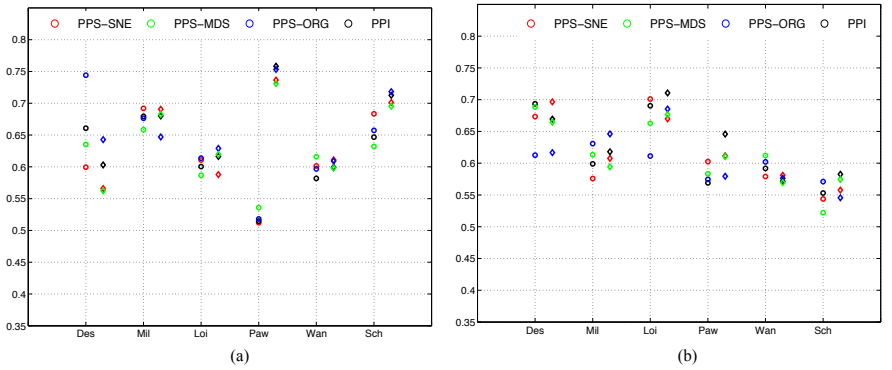

Fig. 4. Classification performance (AUC) of using four different protein association terms in determining subnetwork matrix per dataset. a) early and b) late integration approach. "circle" refers to the geneset passing and "diamond" refers to the classifier passing strategy.

3.2 Functionally Coherent Subnetworks

Table 2 shows that generally, the size of the networks obtained by $\boldsymbol{PPS}$ through all datasets is smaller than the obtained networks using $\boldsymbol{PPI}$. This is especially apparent for the results with $\boldsymbol{PPS}_{ORG}$. The significant subnetworks obtained with $\boldsymbol{PPS}_{SNE}$ as well as with $\boldsymbol{PPS}_{MDS}$ contain, on average, longer paths and nodes of lower degree than $\boldsymbol{PPS}_{ORG}$ (Fig. 5). However, the overlap with $\boldsymbol{PPI}$ based subnetworks, in terms of genes, is substantially higher than $\boldsymbol{PPS}_{ORG}$ (32% compared to 47%, Fig. 6). Noteworthy, in all approaches, genes within the significant subnetworks almost exclusively consist of genes that are either positively or negatively associated with the prognosis labels of the samples (Fig. 5).

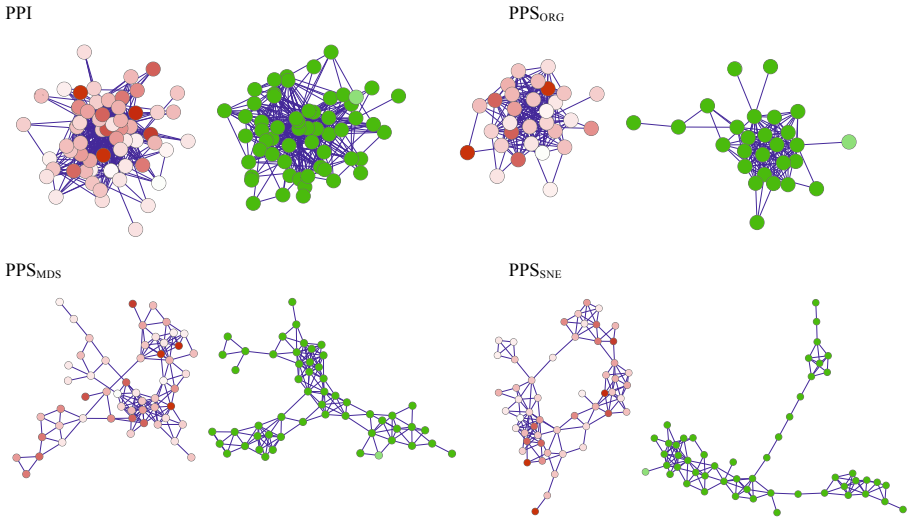

Fig. 5. The two largest significant subnetworks using the early integration approach for four different protein association matrices ($\boldsymbol{P}_B$) in $\boldsymbol{S}$. Genes are colored based on the p-value of the Welch t-test for their association with the Poor and Good expression outcome. Red and green indicate higher expressed in Good and Poor, respectively.

To find out whether the function of discriminative genes are enriched for cancer-related functional categories, the DAVID tool is applied [21]. This enrichment analysis reveals that for all four protein association matrices $\boldsymbol{P}_B$ enrichment for hallmark cancer categories is observed. However, this enrichment is substantially stronger for the $\boldsymbol{PPS}_{SNE}$ and $\boldsymbol{PPS}_{MDS}$ methods as indicated in Table 3. Comparing the genes in the subnetworks resulting from the $\boldsymbol{PPS}_{SNE}$, and $\boldsymbol{PPI}$ matrices reveals that genes within the large subnetworks of $\boldsymbol{PPS}$ strongly overlap with $\boldsymbol{PPI}$ (Fig. 7). To find out if one of the two methods is better able to capture biological relevant genes (i.e. genes with a relation to cancer), we performed a functional comparison and found that, the common core of

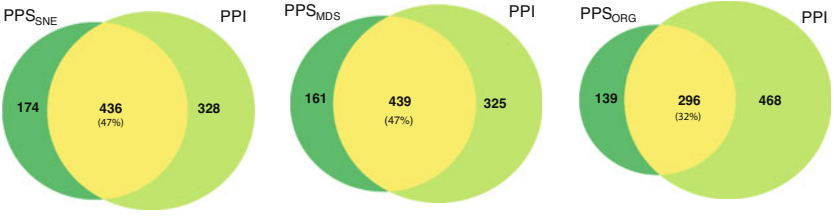

Fig. 6. Venn diagrams representing overlaps in number of the selected genes in significant subnetworks obtained by $\boldsymbol{PPI}$ and $\boldsymbol{PPS}$s using early integration approach. The mapped $\boldsymbol{PPS}$ share more genes with $\boldsymbol{PPI}$.

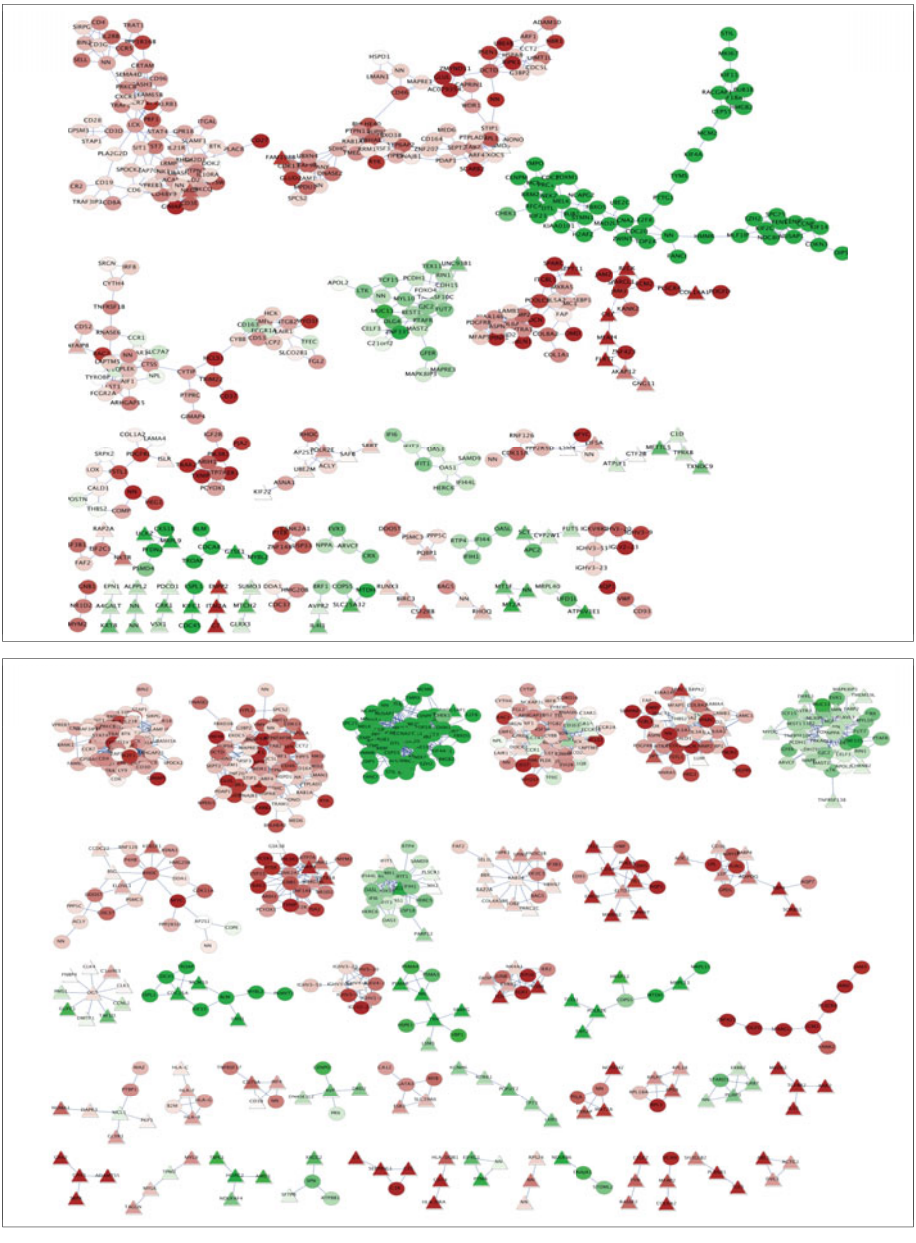

Fig. 7. Significant subnetworks ($T_S < 0.05$) with minimal size of three genes using early integration approach by a) $\boldsymbol{PPS}_{SNE}$, b) $\boldsymbol{PPI}$. "Circle" nodes indicate the overlapping genes between two methods ($\boldsymbol{PPS}_{SNE}$ and $\boldsymbol{PPI}$) and "Triangular" nodes refer to gene that are exclusive for one of the methods. Genes are colored based on the p-value of the Welch t-test for their association with the Poor and Good expression outcome. Red and green indicate higher expressed in Good and Poor, respectively.

the genes in these subnetworks are significantly enriched for the hallmark process of breast cancer (Table 4) e.g., mitotic cell cycle (GO:0000278, p-value: 6.4e-18) nuclear division (GO:0000280, p-value:2.3e-16), mitosis (GO:0007067, p-value: 2.2e-16).

Most striking is the most significant GO category found for the gene set exclusive for $\boldsymbol{PPS}_{SNE}$ (GO:0070013, intracellular organelle lumen) which contains the highly relevant breast cancer associated BRCA1 and BRCA2 genes. This demonstrates the efficacy of the $\boldsymbol{PPS}_{SNE}$ to detect genes that have direct causal implications in breast cancer. In addition, the significant gene sets are analyzed using IPA (Ingenuity Systems) [22] to explore the genes strongly associated with cancer. The functional analysis identifies putative biomarkers of breast cancer process selected exclusively by $\boldsymbol{PPS}_{SNE}$ such as FAM198B (Entrez gene ID: 51313) [23], KIF22 (Entrez gene ID: 3835) [24], [25] and FLRT2 (Entrez gene ID: 23768) [26].

Table 3. Gene ontology enrichment and their approximated $\log(p-value)$ associated with the significant subnetworks in early integration method

GO Terms	$\boldsymbol{PPI}$	$\boldsymbol{PPS}_{ORG}$	$\boldsymbol{PPS}_{MDS}$	$\boldsymbol{PPS}_{SNE}$
GO:0000278 mitotic cell cycle	**20**	8	18	**20**
GO:0048285 organelle fission	13	4	16	**18**
GO:0000280 nuclear division	14	4	17	**18**
GO:0007067 mitosis	14	4	17	18
GO:0000087 M phase of mitotic cell cycle	14	4	16	**18**
GO:0000279 M phase	14	5	15	**16**
GO:0022403 cell cycle phase	13	4	14	**15**
GO:0007049 cell cycle	**19**	7	14	13
GO:0022402 cell cycle process	**18**	6	15	14
GO:0051301 cell division	9	5	11	**13**
GO:0007059 Chromosome segregation	6	-	9	**10**
GO:0005819 spindle	**10**	2	8	9

3.3 Consistent Consensus Genes across the Datasets

To investigate the consistency in gene selection across the six different expression datasets, we analyze the similarities and diversities of the significant subnetworks derived by the four alternatives of the $\boldsymbol{P}_B$ matrix in $\boldsymbol{S}$. Pairwise similarities are computed by applying two criteria, Jaccard index and odds ratio. The Jaccard index obtained by employing mapped $\boldsymbol{PPS}$ (on average 20%) surpasses the other methods. This means that, when $\boldsymbol{PPS}_{SNE}$ or $\boldsymbol{PPS}_{MDS}$ is used to detect significant subnetworks, the signatures are more concordant across datasets (Table 5).

Table 4. Top five enriched GO FAT terms for selected genes within significant subnetworks with minimal size of three genes i) exclusively by $\boldsymbol{PPS}_{SNE}$ (dark green in Fig. 6), ii) shared by ***PPI*** and $\boldsymbol{PPS}_{SNE}$ (yellow in Fig. 6) iii)exclusively by ***PPI*** (light green in Fig. 6)

$\boldsymbol{PPS}_{SNE}$	Common	***PPI***
GO:0070013	GO:0000278	GO:0010033
intracellular organelle lumen	mitotic cell cycle	response to organic substance
GO:0043233	GO:0000280	GO:0009725
organelle lumen	nuclear division	response to hormone stimulus
GO:0031974	GO:0007067	GO:0009719
membrane-enclosed	mitosis	response to endogenous stimulus
GO:0044420	GO:0000087	GO:0048545
extracellular matrix	M phase of mitotic cell cycle	response to steroid stimulus
GO:0031981	GO:0048285	GO:0032570
nuclear lumen	organelle fission	response to progesterone stimulus

Table 5. Pairwise comparison of the significant networks obtained using four protein association matrices on six datasets, the mean of the Jaccard index and odds ratio

	PPI	$\boldsymbol{PPS}_{ORG}$	$\boldsymbol{PPS}_{MDS}$	$\boldsymbol{PPS}_{SNE}$
Jaccard Index	0.16	0.13	0.2	0.2
Odds Ratio	20.5	29.1	30.2	28.6

The odds ratio confirms this conclusion. The significant subnetworks selected by $\boldsymbol{PPS}_{MDS}$ (30.2%) and $\boldsymbol{PPS}_{SNE}$ (28.6%) demonstrate the ability of selecting consistent predictive genes across the different expression datasets.

4 Conclusion

We provide a novel protein similarity measure based on the protein family information contained in Pfam and use this to select discriminative markers on six breast cancer gene expression datasets. This protein similarity matrix captures functional information by focusing on shared and evolutionary conserved protein domains and other sequence similarities. We have demonstrated that the obtained significant genes exhibit more concordance across the six expression datasets. In particular, using the SNE approach to reduce dimensionality of the similarity matrix results in a promising coherence in the selected signature genes across the different datasets. The GO enrichment analysis indicates that the genes that are found for both PPI as well as PPS methods have strong links with cancer. Most importantly, however, for the genes found exclusively using the PPS information substantial evidence is available to link them to breast cancer.

The proposed method to infer protein similarities results in a promising data source to take into account while searching for marker genes and networks associated with metastasis. Since it only relies on protein sequence and Pfam information a much larger part of the protein space can be included in the marker discovery process. Therefore it is envisioned that these results will also be useful in other molecular classification problems.

Acknowledgments. This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by the Netherlands Genomics Initiative (NGI).

References

1. Weigelt, B., et al.: Breast cancer metastasis: markers and models. Nat. Rev. Cancer 5(8), 591–602 (2005)
2. Veer, L.J., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871), 530–536 (2002)
3. Vijver, M.J., et al.: A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. 347(25), 1999–2009 (2002)
4. van Vliet, M.H., et al.: Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. BMC Genomics 9, 375 (2008)
5. Ein-Dor, L., et al.: Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21(2), 171–178 (2005)
6. Hua, J., Tembe, W.D.: Performance of feature-selection methods in the classification of high-dimension data. Pattern Recog. 42(3), 409–424 (2009)
7. Symmans, W.F., et al.: Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions. Hum. Pathol. 26(2), 210–216 (1995)
8. Shen, R., et al.: Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC Genomics 5(1), 94 (2004)
9. Pujana, M.A., et al.: Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat. Genet. 39(11), 1338–1349 (2007)
10. Chuang, H.Y., et al.: Network-based classification of breast cancer metastasis. Mol. Sys. Bio. 3, 140 (2007)
11. van den Akker, E., et al.: Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis (submitted)
12. Rigden, D.: From protein structure to function with bioinformatics. Springer, Heidelberg (2009)
13. Finn, R.D., et al.: The Pfam protein families database. Nucleic Acids Res. 38, D211–D222 (2010)
14. Eddy, S.R.: A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comp. Bio. 4(5), e1000069 (2008)
15. von Mering, C., et al.: STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 31(1), 258–261 (2003)
16. Biomart, `http://www.biomart.org/biomart/martviewrt`
17. van der Maaten, L.J.P., Hinton, G.E.: Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Res. 9, 2579–2605 (2008)
18. Goeman, J.J., et al.: A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20(1), 93–99 (2004)
19. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et des Jura, Bulletin de la Société Vaudoise de Sciences. Naturelles 37, 547–579 (1901)
20. Edwards, A.W.F.: The measure of association in a 2×2 table. JSTOR 126(1), 1–28 (1968)

21. Huang, D.W., et al.: Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 4(1), 44–57 (2009)
22. Ingenuity Pathways Analysis software, `http://www.ingenuity.com`
23. Deblois, G., et al.: Genome-wide identification of direct target genes implicates estrogen-related receptor alpha as a determinant of breast cancer heterogeneity. Cancer Res. 69(15), 6149–6157 (2009)
24. Yumei, F.: KNSL4 is a novel molecular marker for diagnosis and prognosis of breast cancer. American Assoc. for Cancer Res. (AACR) Meeting Abstracts, 1809 (2008)
25. Diarra-Mehrpour, M., et al.: Prion protein prevents human breast carcinoma cell line from tumor necrosis factor alpha-induced cell death. Cancer Res. 64(2), 719–727 (2004)
26. Tripathi, A., et al.: Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. Int. J. Cancer 122(7), 1557–1566 (2008)