Chapter 4

Predicting Cancer-Related Proteins Using Network Contextual Information

Based on

Hossein Rahmani, Hendrik Blockeel and Andreas Bender, "Predicting Genes Involved in Human Cancer Using Network Contextual Information" Journal of Integrative Bioinformatics, 9(1):210, 2012.

4.1 Abstract

Protein-Protein Interaction (PPI) networks have been widely used for the task of predicting proteins involved in cancer. Previous research has shown that functional information about the protein for which a prediction is made, proximity to specific other proteins in the PPI network, as well as local network structure are informative features in this respect. In this work, we introduce two new types of input features, reflecting additional information: (1) Functional Context: the functions of proteins interacting with the target protein (rather than the protein itself); and (2) Structural Context: the relative position of the target protein with respect to specific other proteins selected according to a novel ANOVA (analysis of variance) based measure. We also introduce a selection strategy to pinpoint the most informative features. Results show that the proposed feature types and feature selection strategy yield informative features. A standard machine learning method (Naive Bayes) that uses the features proposed here outperforms the current state-of-the-art methods by more than 5% with respect to F-measure. In addition, manual inspection confirms the biological relevance of the top-ranked features.

4.2 Introduction

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks [118]. Much can be learned from the analysis of such networks with respect to the metabolic and signalling processes present in an organism, and the knowledge gained can also be prospectively employed e.g. to the task of protein function prediction [78, 98, 111, 121, 119, 57, 13, 18], identification of functional modules [71], interaction prediction [48, 129], identification of disease candidate genes [106, 37, 132, 87] and drug targets [104, 81], according to an analysis of the resulting network [72].

Wu et al. [130] present an excellent overview of multiple methods for detecting proteins involved in cancer or disease. Among the different methods discussed in [130], "guilt-by-proximity" methods are well known. Methods classified in this category are based on the assumption that genes that directly interact, or, more generally, lie close to each other in the network, are more likely to be involved in the same diseases (as argued by, e.g., Gandhi et al. [31]). The methods vary based on how they define proximity: Some methods consider only direct neighbors to be in the proximity (e.g., [87, 3]), some quantify proximity of two proteins using the length of the shortest-path between them, some compute a "Global Distance Measure" that also takes into account how many paths there are between the two proteins, and how long these are; an example is the approach by Chen et al. [16], who use a PageRank based model for this.

While the basic guilt-by-proximity methods require that certain nodes in the network are already known to be involved in the disease under study, Wu et al. also discuss methods that rely on proximity to nodes known to be involved in other, simi-

lar diseases. Wu et al. define *de novo* methods as methods that can predict nodes to be involved in a particular disease even if no other nodes in the network are known to be involved in it.

The methods discussed by Wu et al. mostly rely on notions of proximity (to genes known to be disease-related) from the area of graph analysis. An entirely different type of approaches are those that rely on feature-based descriptions. There, each individual protein is described by means of a fixed set of features. Next, using machine learning methods, a model is learned that links some of these features to disease-relatedness. In the context of predicting involvement in cancer, examples of feature-based methods include Milenkovic et al.[77], Furney et al. [29] and Li et al. [66]. Milenkovic et al. [77] characterize a protein using a "signature vector" that describes the local network structure around the node in terms of so-called graphlets, small fixed graph structures in which the node occurs. By applying a series of clustering methods, they show that protein that are involved in cancer have similar "topological signatures", which distinguish them from other proteins, and these nodes need not be close to each other in the network. Furney et al. [29] use the Gene Ontology annotations of a protein as features, as well as a number of other properties; they use a chi-square-based selection criterion to select the likely most relevant features, then apply Naive Bayes. Li et al. [66] compare three classifiers: SVM, Naive Bayes and logistic regression and they find that the SVM classifier on average performed slightly better than the Naive Bayes and logistic regression methods, and that among SVMs using different types of features individually, including GO annotations as features gives the best performance, while sequence and conservation features have relatively weak predictive power.

When learning from PPI networks, feature-based approaches have a number of advantages over proximity-based approaches. First, defining the problem in a machine learning setting gives access to a wide range of machine learning techniques, making this type of approaches very flexible. Second, data integration is more easily achieved: one can easily define additional features for proteins, possibly using background information (i.e., information external to the PPI network) for this. Third, these method are inductive: they do not yield predictions, but a model for making predictions. This is interesting in terms of Wu et al.'s definition of *de novo* methods. Information about disease genes is needed when *constructing* the model, but not when *applying* it, so the model can be applied to other PPI networks, or in other areas of the same PPI network. Finally, inductive methods can yield interpretable models, which may by themselves yield new insights.

A difficulty with feature-based methods, however, is that the quality of the learned model depends on the features used. When the input data is a PPI network, the main challenge is to find features with good predictive power that can be computed from this network. The approaches mentioned above all do this in some way. In this work, we propose two new types of input features, reflecting additional information that can be extracted from a PPI network: (1) Functional Context: the functions of proteins interacting with the target protein (rather than the protein itself); (2) Structural Context: the relative position of the target protein with respect to specific other proteins selected according to a novel ANOVA (analysis of variance) based

measure. We show that these features have good predictive power. It is not our goal to compare different machine learning algorithms; we restrict ourselves to the Naive Bayes classifier. The performance of the method might be optimized by using another learning method, but we expect the difference to be small (see also Li et al. [66]). Our main claim lies in the usefulness of the new features.

4.3 Methods

4.3.1 Formal Definition

We consider a PPI network as an undirected annotated graph (P, E, λ) where P is a set of proteins, $E \subseteq P \times P$ is a set of interactions between these proteins, and λ is a so-called annotation function; for each p, $\lambda(p)$ denotes the additional information we have about p (for instance, its GO annotations). In this work, we assume that $\lambda(p)$ simply lists all the GO functions that are associated with p; we call it the function set (or function vector) of p, and denote it $FS(p)$. If $F = \{f_1, f_2, \ldots, f_n\}$ is the set of all the functions in the network, then $FS(p)$ is an $|F|$-dimensional binary vector; the i'th component of $FS(p)$, denoted $FS_i(p)$, is 1 if function f_i is associated with p, and 0 otherwise. We will also write $f_i \in FS(p)$ to denote $FS_i(p) = 1$.

4.3.2 Protein Description Based on Functional Context

Given a protein p, we define the interactor set of p, denoted $IS(p)$, as the set of proteins it interacts with, i.e., $IS(p) = \{q|(p, q) \in E\}$. Besides the function vector of p itself, we also define the "interacting function counts" vector $IFC(p)$ as the number of interacting proteins that are annotated with that function.

$$IFC(p) = \sum_{q \in IS(p)} FS(q) \tag{4.1}$$

Note that, while methods for predicting involvement in cancer have considered GO annotations of proteins as predictive features (e.g., [29, 66]), no methods up till now have considered GO annotations of the neighbors of those proteins at the same time. That is, for predicting involvement in cancer of a protein p, the $FS(p)$ vector has been considered as a predictive feature, but the vector $IFC(p)$ has not. One may wonder what the advantage is of using GO annotations of related proteins, rather than the protein itself. One argument is that GO annotations are often incomplete, and by collecting GO information from the neighbors of a protein p, we may get more information about p itself. This argument is backed up by the fact that GO annotations of proteins can often be predicted well from the GO annotations of their neighbors; see, e.g., [111, 99]. However, as we will show, this is not the only effect; there is also a direct relationship between a protein's involvement in cancer and the GO annotations of the proteins it interacts with.

We will refer to the information in $FS(p)$ and $IFC(p)$ as the *functional context* of p. We experimentally compare two different versions of this functional context: using $FS(p)$ only as input vector (i.e., ignoring the information in the neighborhood of p), and using the sum of $FS(p)$ and $IFC(p)$ as input vector (thus taking into account functional information about the neighborhood of p, including p itself). We call these two approaches FS and $FS + IFC$, respectively.

As defined above, the $FS(p)$ and $IFC(p)$ vectors have high dimensionality; the number of components equals the number of functions in the Gene Ontology. A natural way to reduce this dimensionality is using a feature selection method to filter out the least interesting features (functions, in this case). An often used measure for determining the relevance of a binary feature F for a class variable C is the χ^2 score, defined as follows:

$$\chi^2 = \frac{(ad - bc)^2 * (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \tag{4.2}$$

where a, b, c and d are defined by the contingency table in Table 4.1.

Table 4.1: **The contingency table of a binary feature F w.r.t. a binary class variable C**

	$F = 0$	$F = 1$	total
C=0	a	b	a+b
C=1	c	d	c+d
	a+c	b+d	a+b+c+d

a, b, c, and d count the number of times F and C have the corresponding value. The χ^2 value of F w.r.t. C is derived from this.

In our case, the class variable C indicates whether a protein p is involved in cancer or not, and the binary feature F indicates whether a particular component of $FS(p)$ or $FS(p) + IFC(p)$ is zero ($F = 0$) or not ($F = 1$).

Apart from allowing us to reduce the dimensionality of the vectors describing a protein p, the χ^2 measure also ranks functions from highly relevant (for predicting involvement in cancer) to less relevant.

4.3.3 Protein Description Based on Structural Context

Besides the functional context of a protein, defined before, we will also consider its so-called *structural context*. This structural context relates to the relative position of p in the network.

Several methods discussed in Wu et al. [130] describe each protein p based on the shortest-path distance of p to some previously known cancer/disease proteins. We refer to this category of methods as "distanceToCancer" methods (DisToCancer).

Alternatively, we can describe a protein's position relative to other proteins than only cancer-related ones. Rahmani et al. [98] proposed a relevance measure for

proteins that is inspired by statistical ANOVA (analysis of variance), and showed that shortest-path distance to a relatively small number of proteins (selected according to the ANOVA-based measure) is informative for the task of function prediction in the PPI networks. Since the ANOVA method works well for function prediction, it is natural to check whether it also gives good results for the task of predicting cancer-related proteins, and this is one of the purposes of the current study. We therefore propose the use of similar features for predicting proteins involved in cancer.

The ANOVA-inspired selection measure (briefly, ANOVA) is defined as follows. Let P^+ be the set of proteins labeled as being involved in cancer, and P^- the set of proteins not labeled as such. For each protein q, we introduce a feature d_q; $d_q(p)$ denotes the shortest-path distance between p and q (viewed here as a feature of p). We consider for each q the mean and variance of $d_q(p)$, taken over all cancer-related and non-cancer-related p:

$$m_q^+ = \frac{\sum_{p \in P^+} d_q(p)}{|P^+|} \tag{4.3}$$

$$m_q^- = \frac{\sum_{p \in P^-} d_q(p)}{|P^-|} \tag{4.4}$$

$$var_q^+ = \frac{\sum_{p \in P^+} (d_q(p) - m_q^+)^2}{|P^+| - 1} \tag{4.5}$$

$$var_q^- = \frac{\sum_{p \in P^-} (d_q(p) - m_q^-)^2}{|P^-| - 1} \tag{4.6}$$

Seeing P^+ and P^- as two groups of proteins, the following formula compares the variance between groups to the variance within groups (as it is used for relative ranking only, constant factors are dropped):

$$A_q = \frac{(m_q^+ - m_q^-)^2}{var_q^+ + var_q^-} \tag{4.7}$$

A high A_q means that d_q varies little within groups and/or much between groups, which indicates that d_q has high predictive power for the group. Features d_q can be ranked according to A_q, and the top-k features selected as actual features to be included in the description of all proteins. We will call the category of methods that use these features DisToAnova methods, or DisToAnova(k) when referring to a particular setting for the parameter k.

Finally, we can combine the information in the DisToCancer and DisToAnova descriptors; we do this by first filtering the proteins, retaining only those known to be involved in cancer, and ranking these according to the Anova criterion. This combined version is referred to as DisToCancerAnova.

4.3.4 Protein Description Based on Functional and Structural Context

This refers to protein descriptions that include both information from functional and structural context. The input consists of the $FS + IFC$ vector concatenated with the DisTo(Cancer/Anova/CancerAnova) vector.

4.4 Results

4.4.1 Dataset

We evaluate our methods on the dataset used by Milenkovic et al. [77]. This dataset is the union of three human PPI datasets: HPRD [91], BIOGRID [116] and the dataset used by Radivojac et al. [97]. When we say "union", we mean that the new network contains all the nodes and edges (proteins and interactions) found in either of these networks. The aim of merging these three datasets was to obtain as complete a human PPI network as possible, i.e., a network that covers with its edges as many proteins in the human proteome as possible. We denote as "known cancer genes" the set of genes implicated in cancer that is available from the following databases: Cancer Gene Database [23], Cancer Genome Project-the Cancer Gene Census [95], GeneCards [32] Kyoto Encyclopedia of Genes and Genomes [83] and Online Mendelian Inheritance in Man [47]. Some statistical information is shown in Table 4.2. We have chosen to evaluate our methods on this dataset to make a precise quantitative comparison to their graphlet-based method possible.

Table 4.2: Statistical information of union of three human PPI datasets: HPRD [91], BIOGRID [116] and Radivojac et al. [97]

Number of Proteins	10,282
Average Degree	9.201
Min Degree	1
Max Degree	272
Number of Cancer Genes	939

While the dataset employed here is of high quality, as it is based on large and widely employed datasets, it should be kept in mind that it is not trivial (or in the narrow sense probably even impossible) to define it in a flawless fashion. One of the limitations lies in the role of 'genes involved in cancer' - cancers are different, so while a gene may play a role in one cancer, it might play no role at all in another one. Also, there are spatial and temporal conditions involved in the annotation we do not include here. On the other hand, a limitation lies in the construction of the interactome we define in our dataset. Again, temporal conditions are excluded, and likely many interactions have not been identified in experiment yet; hence our dataset

likely contains a substantial number of missing annotations (while likely also false positive interactions are included due to experimental noise and errors). Nonetheless, the dataset employed here is as good as we can do currently both in size and quality; and, in particular, it has been employed in related studies before, which enables us to perform benchmark experiments in a comparative manner by utilizing it.

This dataset determines uniquely the network structure, and therefore the values of all features, used in our experiments. The actual datasets we use differ only with respect to what features are included.

4.4.2 Biological Interpretation of the Most Relevant Functions

The number of different functions occurring in our human dataset is 9833; this is also the dimensionality of FS and IFC if no dimensionality reduction is used. As mentioned before, we can use a χ^2-based feature selection method to reduce this number; at the same time, this technique ranks functions according to how relevant they are for prediction of cancer involvement.

Most Relevant Functions in FS

Tables 4.3, 4.4 and 4.5 show the 20 highest ranked functions. As the Gene Ontology actually uses three domains (biological function, molecular function, cellular component), we have separated the functions according to their domain.

Searching the most relevant functions in the cancer literature proved the usefulness of chi-square for detecting these functions. For example, based on cancer literature, function "GO:0008284" is involved in various cancers: "Breast Cancer", "Prostate Cancer" and "Lung Cancer". Besides using the statistic to select a limited number of features, we can also use it to inspect the top-ranked functions, which can be used both as a soundness check (are the functions that we expect to be relevant indeed highly ranked?) and as a method for discovering potentially new information (when there are unexpected functions among the top-ranked ones).

Many of the biological functions contained in Table 4.3 are obviously related to cancer, such as GO:0008284, the Positive regulation of cell proliferation, which is a synonym for uncontrolled cell growth, as are positions 3, 5 and 10 in the list (GO:0045944 Positive regulation of transcription from RNA polymerase, GO:0006355 Regulation of transcription, DNA-dependent and GO:0045941 Positive regulation of transcription). Similarly, position 4 (GO:0008285 Negative regulation of cell proliferation) has an obvious connection to cancer; where positive stimulation of cell growth can stimulate tumor growth, an inhibition of the negative regulatory elements will have the very same effect. Fibroblasts are involved in wound healing, a process not taking place properly in cancerous settings [50]. We can also find biological processes linked to small molecules in the list, at positions 6 and 7, namely GO:0014070 Response to organic cyclic substance and GO:0042493 Response to drug. It is known that many carcinogenic substances such as benzo[a]pyren, or even smaller molecules

Table 4.3: **Most discriminative functions from Biological Process based on *FS* method**

Index	Function	Short Info	chi-square	p-value
1	GO:0008284	Positive regulation of cell proliferation	163.02	<0.0001
2	GO:0008543	Fibroblast growth factor receptor signaling pathway	105.80	<0.0001
3	GO:0045944	Positive regulation of transcription from RNA polymerase II promote	99.65	<0.0001
4	GO:0008285	Negative regulation of cell proliferation	71.40	<0.0001
5	GO:0006355	Regulation of transcription, DNA-dependent	69.84	<0.0001
6	GO:0014070	Response to organic cyclic compound	69.76	<0.0001
7	GO:0042493	Response to drug	69.18	<0.0001
8	GO:0043434	Response to peptide hormone stimulus	67.09	<0.0001
9	GO:0001658	Branching involved in ureteric bud morphogenesis	64.91	<0.0001
10	GO:0045941	Positive regulation of transcription	64.89	<0.0001
11	GO:0007050	Cell cycle arrest	62.73	<0.0001
12	GO:0001656	Metanephros development	62.07	<0.0001
13	GO:0032355	Response to estradiol stimulus	59.99	<0.0001

Table 4.4: **Most discriminative functions from Molecular Function based on *FS* method**

Index	Function	Short Info	chi-square	p-value
1	GO:0016563	Transcription activator activity	88.85	<0.0001
2	GO:0004713	Protein tyrosine kinase activity	84.60	<0.0001
3	GO:0003700	Sequence-specific DNA binding transcription factor activity	83.11	<0.0001
4	GO:0005515	Protein binding	82.88	<0.0001
5	GO:0004716	Receptor signaling protein tyrosine kinase activity	68.76	<0.0001
6	GO:0043565	Sequence-specific DNA binding	67.69	<0.0001

such as benzene, are linked to cancer risk. Unfortunately, one of the limitations of the GO terms is their low selectivity; hence the term 'response to drug' remains rather vague. Positions 9 and 12, GO:0001658 Branching involved in ureteric bud morphogenesis and GO:0001656 Metanephros development, are both linked to growth factors, and hence in turn to the development of cancers.

Molecular functions returned as significantly enriched among cancer genes, listed

Table 4.5: **Most discriminative function from Cellular Component based on FS method**

Index	Function	Short Info	chi-square	p-value
1	GO:0005634	A membrane-bounded organelle of eukaryotic cells in which chromosomesare housed and replicated.	99.76	<0.0001

in Table 4.4, frequently refer to transcription factor (position 1) and kinase activity (positions 2 and 5). On the other hand, the cellular component category was less revealing, only listing one significantly enriched category related to cancer genes - the nucleus (where increased transcription takes place, leading to uncontrolled cell growth). Unfortunately, the GO term employed is too general to draw more detailed conclusions from this analysis.

Most Relevant Functions in $FS + IFC$

18 out of 20 functions with the highest χ^2, calculated based on the $FS+IFC$ method, belong to the Biological Process ontology and are listed in Table 4.6. As is apparent from Table 4.6 (when compared to Table 4.3, which results from the use of the FS method), very different discriminative GO terms from the Biological Process ontology are retrieved. Many biological processes retrieved by this method seem to be more specific, such as GO:0043491 at position 1, naming the protein kinase B signaling cascade as involved in cancerogenesis (which is known from literature [12]). It is interesting that now also secondary processes known to be relevant for cancerogenesis and, in particular, cancer growth and the formation of metastases, are captured (which was not the case by purely applying the FS method), such as at position 6 (GO:0001525) for the formation of blood vessels essential for the rapid growth of cancerogenous tissue, and at position 12 (GO:0030335) with respect to cell migration, important for the formation of metastases. Also novel in the list are biological processes related to insulin and the insulin-like growth factor receptor (IGFR), at positions 2 (GO:0048009) and 5 (GO:0032869). This is supported by literature, as insulin has been linked to pancreatic cancer development [28], while the literature regarding insulin-like growth factor receptor is still inconclusive [19, 88]. Still, due to their apparent role in cell proliferation, it is certainly a possibilty that IGFRs play a role in the development of at least some cancer subtypes.

As shown in Table 4.6, chi-square values when calculated based on $FS + IFC$ are greater than the chi-square values when we use the FS method for the calculation, illustrating how our additional annotations add information to the feature selection step; P-values of all the highly ranked functions are < 0.0001 which is very significant.

Overall, from the discussion above, it becomes apparent that the $FS + IFC$ method, as proposed in this work, is able to retrieve significantly different biologi-

cal processes, compared to using the FS method; thus it adds to the information that can be gained from the same set of data. Hence, we suggest it to be a worthwhile method to be employed in the analysis of signaling networks, as shown in this particular case study.

Table 4.6: **Most discriminative functions from Biological Process based on $FS + IFC$ method**

Index	Function	Short Info	chi-square	p-value
1	GO:0043491	Protein kinase B signaling cascade	280.70	<0.0001
2	GO:0048009	Insulin-like growth factor receptor signaling pathway	231.72	<0.0001
3	GO:0008284	Positive regulation of cell proliferation	223.62	<0.0001
4	GO:0034097	Response to cytokine stimulus	223.05	<0.0001
5	GO:0032869	Cellular response to insulin stimulus	218.56	<0.0001
6	GO:0001525	Angiogenesis	213.14	<0.0001
7	GO:0043066	Negative regulation of apoptosis	211.77	<0.0001
8	GO:0001701	In utero embryonic development	208.366	<0.0001
9	GO:0009887	Organ morphogenesis	207.71	<0.0001
10	GO:0042493	Response to drug	205.81	<0.0001
11	GO:0030097	Hemopoiesis	202.45	<0.0001
12	GO:0030335	Positive regulation of cell migration	202.38	<0.0001
13	GO:0051091	Positive regulation of sequence-specific DNA binding transcription factor activity	194.37	<0.0001
14	GO:0046326	Positive regulation of glucose import	194.13	<0.0001
15	GO:0043627	Response to estrogen stimulus	192.34	<0.0001
16	GO:0044419	Interspecies interaction between organisms	191.29	<0.0001
17	GO:0014070	Response to organic cyclic compound	189.94	<0.0001
18	GO:0045944	Positive regulation of transcription from RNA polymerase II promoter	189.32	<0.0001

4.4.3 Biological Interpretation of the Most Discriminative Proteins

Our dataset contains 10,282 proteins. The DisToAnova method uses the ANOVA measure to select the most relevant among these. More detailed information could be obtained from an ANOVA analysis of the most relevant proteins among the full set of 10,282 proteins. Table 4.7 shows the 10 proteins with the highest ANOVA measure.

Zinc finger protein (ZNF467) is known to be upregulated in a variety of breast cancers; however usually its close link with BRCA1 has been seen as the reason for its causal relation with cancers [56]. STATIP1 is involved in histone H3 and H4 acetylation and its interactions with STAT3 and JAK1/2 - which are all involved in cell growth and differentiation processes - have been documented in literature [22]. JUNB

has been documented as a proto-oncogene and IL22 (along with its subunit IL22RA2) is involved in Stat3 phosphorylation [15]. FGFR4 (fibroblast growth factor receptor 4) is associated with cancer nearly by definition (and in alignment with fibroblasts being identified earlier in the context of biological functions). The chemokine ligand 1 receptor, CCL1, has been implicated in cancer before and also it has been suggested as a therapeutic target in this context [45]. Platelet derived growth factor C (PDGFC) is part of the PDGFR-alpha signalling pathway and the influence of PDGFR expression on metastatic behaviour has been well documented [126]. STAT1 is involved in cell growth processes [133], hence its appearance in this list is reasonable. C20ORF185 is an interesting case in that it is annotated as possibly being involved in recognizing/binding specific classes of odorants or serving as a defence mechanism by removing pathogenic microorganisms from the mucosa [24]. On the other hand, its recommended name is the "Long palate, lung and nasal epithelium carcinoma-associated protein 3 precursor", rendering its inclusion in the list of proteins most involved in cancer reasonable.

Table 4.7: **Most discriminative proteins based on ANOVA measure**

Index	Protein Name	Official Full Name
1	ZNF467	Zinc finger protein 467
2	STATIP1	Elongator complex protein 2
3	JUNB	Transcription factor jun-B
4	IL22RA2	Interleukin-22 receptor subunit alpha-2
5	FGFR4	Fibroblast growth factor receptor 4
6	CCL1	Cyclin associated with protein kinase Kin28p
7	PDGFC	Platelet-derived growth factor C
8	IL22	Interleukin 22
9	STAT1	Signal transducer and activator of transcription 1-alpha/beta
10	C20ORF185	Long palate, lung and nasal epithelium carcinoma-associated protein 3

4.4.4 Comparing Different Contextual Methods

We divided the dataset into a training set containing 90%, and a test set containing the remaining 10%, of the proteins for the selection of contextual method and tuning. For the final evaluation we use 10-fold cross validation. Features were selected according to the above-mentioned χ^2 and ANOVA methods; in both cases only the training set was used to rank features according to relevance. We have varied the number of features (functions for functional context methods, proteins for structural context methods) from low to high, in order to investigate the effect of this parameter on predictive performance.

With each method, we predict the cancer-relatedness of the nodes in the test set using our various methods, and evaluate the predictions according to Recall, Precision and F-measure:

$$Precision = \frac{tp}{tp + fp} \tag{4.8}$$

$$Recall = \frac{tp}{tp + fn} \tag{4.9}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.10}$$

where proteins involved in cancer are considered as the positive class, and tp, fp and fn denote the number of true positives, false positives, and false negatives, respectively.

Figure 4.1 shows the evaluation metrics for two functional context methods FS and $FS + IFC$ in different function counts: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 with respect to F-measure, Precision and Recall. Independent from the function count, the $FS + IFC$ method always outperforms the FS method with respect to F-measure and this proves our assumption about the usefulness of considering the whole functional context of proteins (not just the functions of the protein itself but also those of its neighbors) for predicting the proteins involved in cancer. The best obtained F-measure with FS is 29% while the best obtained F-measure for $FS + IFC$ is 37% in one case and 35% in three cases.

These results show that considering the functional annotation of the neighbors allows for more accurate prediction of which genes are involved in cancer. Since it was already known that the functional annotation of a protein's neighbors can be used to predict the protein's own functions [111, 99], and that the protein's own functions are relevant for its involvement in cancer [29, 66], one might wonder to what extent our results are simply a consequence of these two facts. We can test this by enriching proteins in the PPI network with predicted GO annotations (predicted from the GO annotations of their neighbors), and next applying the FS method. We tested this by using a Majority Rule method [111] for enriching the GO annotations of the proteins, in two different ways. In the first approach, we perform function prediction for each protein p which $|FS(p)| = 0$ (reasoning that if a protein is not annotated with any functions, it is likely that its functions are simply not known), while in the second, less conservative, approach, we extend the function set of each protein p with $|FS(p)| < 10$ to a total of ten functions. In the notation employed here, the $||$ operator returns the size of the function set of protein p, with 10 being the average function count of proteins in our dataset before applying the Majority Rule method. We call the enriching approaches "Unclassified" and "Extended", respectively. Figure 4.2 compares the FS and $FS + IFC$ methods with their "Unclassified" and "Extended" versions, and we can see that there is no major difference between the original methods employed, compared to their respective functionally enriched versions. This confirms that the functions of neighboring proteins directly influence disease-relatedness; the influence cannot be explained by the relationship between the functions of neighboring

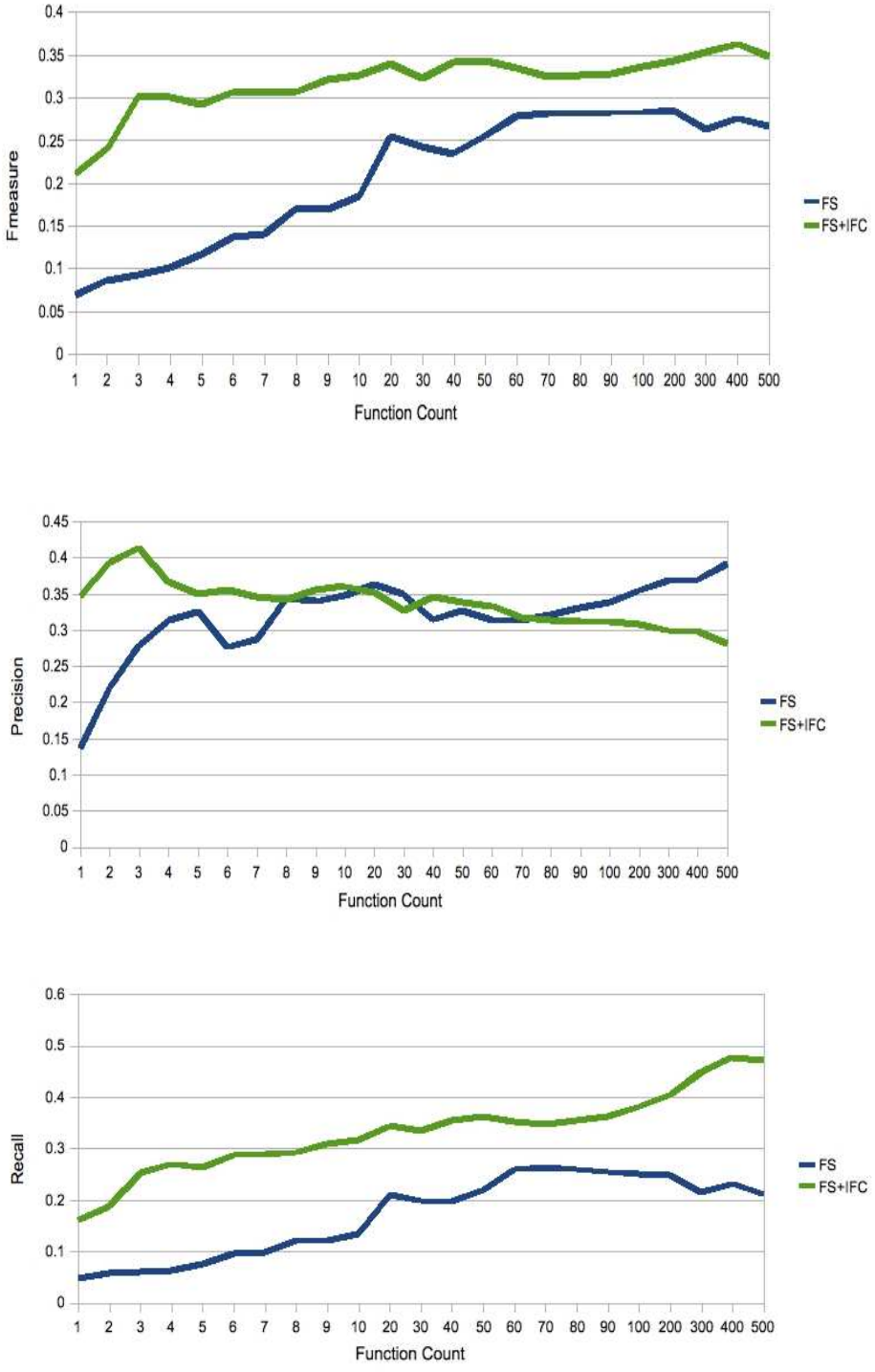

Figure 4.1: Comparing different Functional Context methods. The $FS+IFC$ method always outperforms the FS method with respect to F-measure.

proteins on the one hand, and between a protein's own functions and involvement in disease on the other hand.

Figure 4.3 compares three structural context methods disToCancer, disToAnova and disToCancerAnova in different protein counts: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 with respect to F-measure, Precision and Recall. (As disToCancer has no natural criterion for selecting a subset of cancer-related proteins, proteins were selected randomly in this case, to arrive at comparable counts.) It turns out that, in order to get reasonable F-measure results, selecting less than 30 proteins is enough in the structural context methods. With respect to F-measure, methods using ANOVA for selecting the important proteins almost always outperform the method that selects previously known cancer-related proteins.

In Figure 4.4, we show the result of integrating the functional context method $FS+IFC$ with any one of the three structural context methods, disToAnova, disToCancer and disToCancerAnova. We vary the number of analyzed functions from 5 to 30, and the number of analyzed proteins from 1 to 40. The integration of $FS + IFC$ with disToAnova slightly outperforms the other two integrated methods. Although it may seem that applying the ANOVA method results in only small numerical improvements, Figure 4.4 shows that its integration with the functional annotation of the proteins consistently results in improved results with respect to F-measure values. Compared to functional and structural context methods, the integrated method gives rise to more cases (17 out of 52 in $(FS + IFC)$-DisToCancerAnova, as opposed to 0 out of 52 in $FS + IFC$) with F-measure over 35% (and up to 39% in one case).

4.4.5 Comparing with Previous Methods

Milenkovic et al. [77] have evaluated their method using a leave-one-out cross validation and report an F-measure of 24%. They compare this result to that of Aragues et al. [3], who use information from heterogeneous data sources: (i) Protein Protein Interaction networks, (ii) differential expression data, (iii) structural and functional properties of cancer genes; Aragues et al. report an F-measure of 18.15% for their method. Further, we will compare our results to the method of Furney et al. [29]. As Furney et al. reported results on another dataset, to obtain more comparable results we have implemented their method by selecting 100 functions based on the chi-square value, describing each protein based on those selected functions, and using the Weka machine learning system to apply Naive-Bayes for predicting the proteins involved in cancer.

Our method uses as parameters the number of functions and proteins to be selected by the feature selection method. To optimize these parameters, we divided the human dataset into three parts: 80% for training the model with a particular parameter setting; 10% for tuning the different parameters (that is, models trained with particular parameter values are tested on this 10% and the parameter settings that perform optimally here will be used for the final evaluation), and 10% for evaluating the model; note that this last 10% was not involved in the training in any way. Table 4.8 shows the optimal parameter settings for each method.

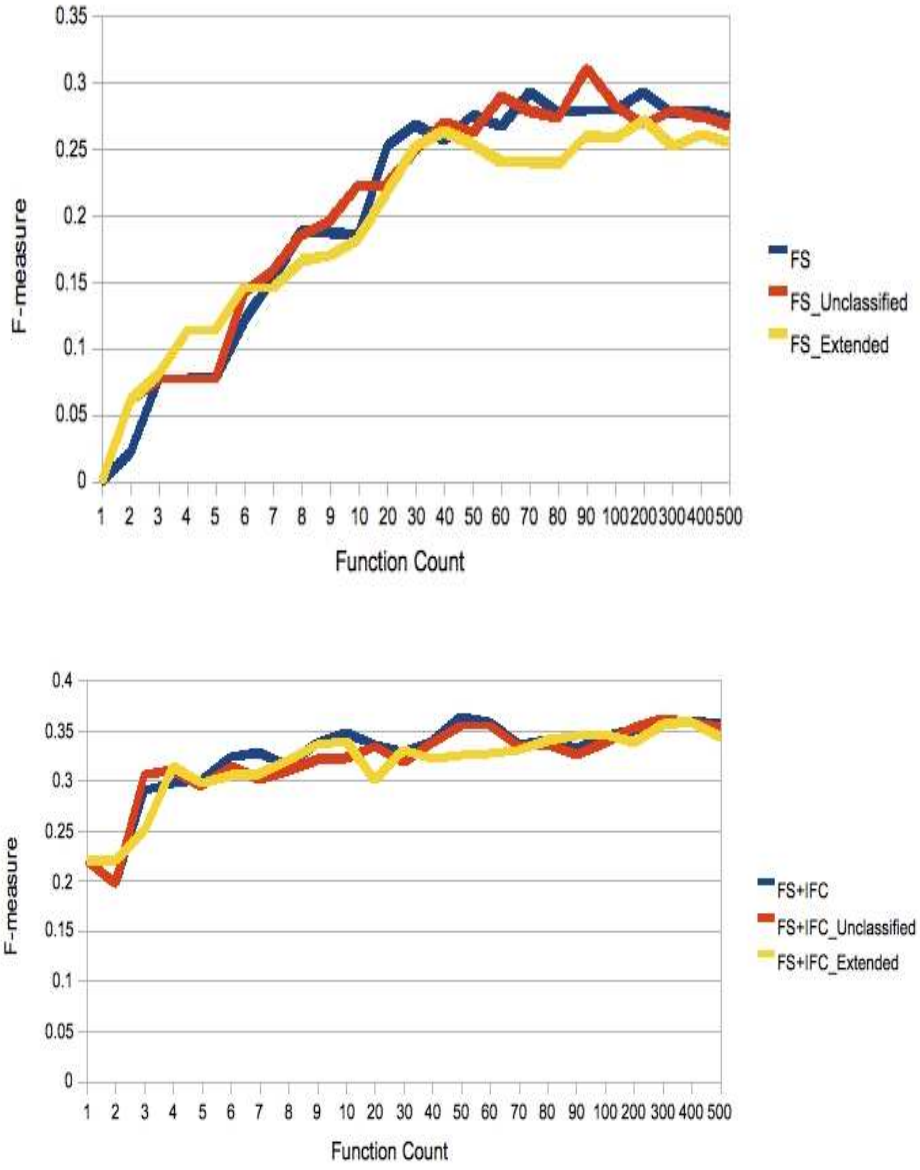

Figure 4.2: Comparing different Functional Context methods with their enriched functional versions. There is no major difference between the original methods employed, compared to their respective functionally enriched versions.

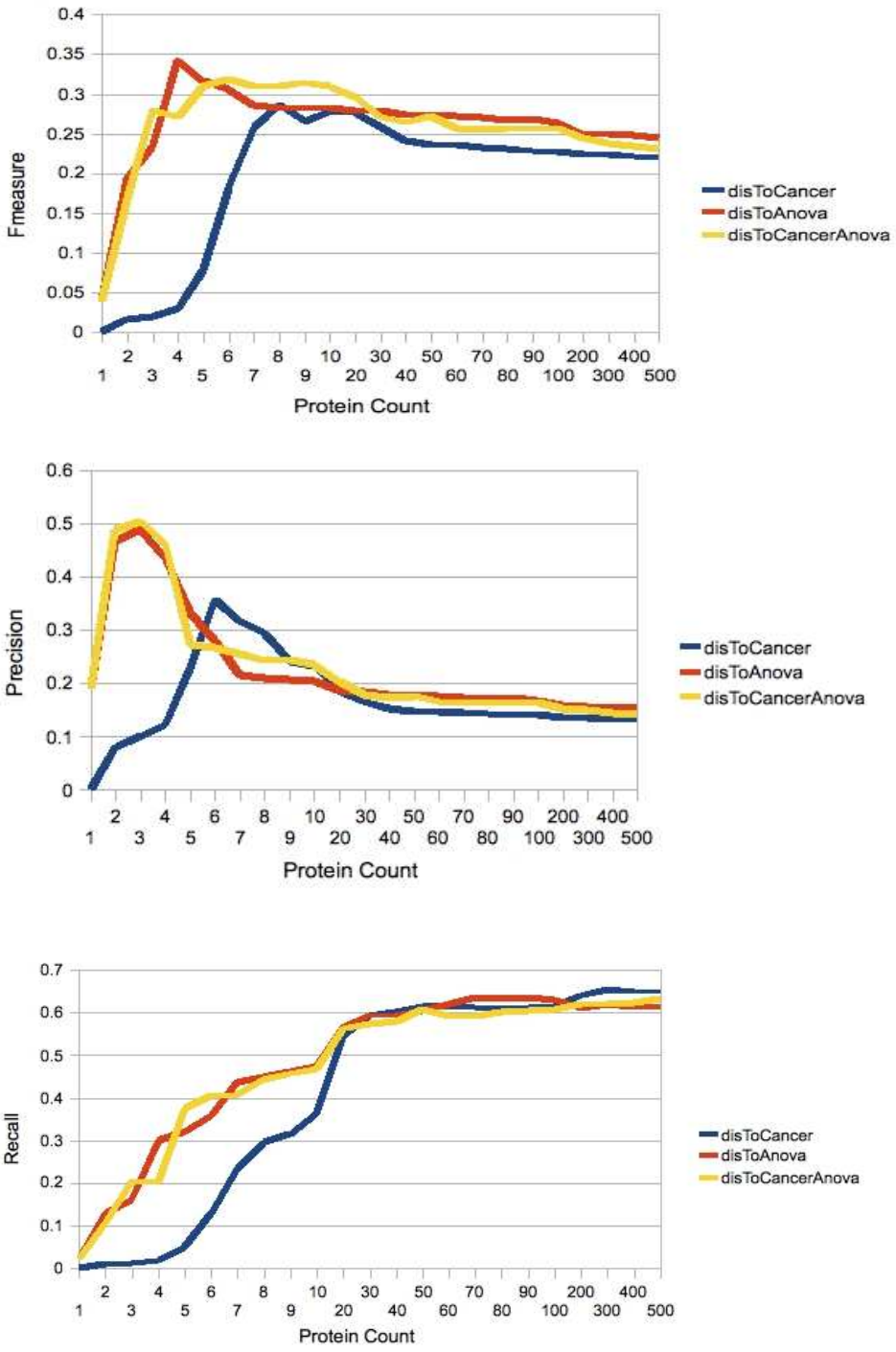

Figure 4.3: Comparing different Structural Context methods. With respect to F-measure, methods using ANOVA for selecting the important proteins almost always outperform the method which selects the previously known cancer-related proteins.

Figure 4.5 compares all the proposed methods with each other using 10-fold cross validation. The method (FS+IFC)-DisToCancerAnova which considers network contextual information outperforms all other proposed methods.

Table 4.8: **Tuning result of each method**

Method Name	Best Feature Count	F-measure in Test Set
FS	90 Functions	28
$FS + IFC$	100 Functions	34
disToAnova	10 Proteins	28
disToCancer	10 Proteins	29
disToCancerAnova	10 Proteins	30
$(FS + IFC)$-DisToCancer	10 Functions and 4 Proteins	35
$(FS + IFC)$-DisToAnova	10 Functions and 5 Proteins	37
$(FS + IFC)$-DisToCancerAnova	10 Functions and 9 Proteins	37

Figure 4.6 compares our best proposed method with previous methods using 10-fold cross validation. The method (FS+IFC)-DisToCancerAnova which considers network contextual information outperforms the previous methods (Furney et al. [29], Aragues et al. [3] and Milenkovic et al. [77]), by 5%, 13% and 8%, respectively, with respect to F-measure.

4.4.6 Random Feature Selection

We showed that the ANOVA method for selection of proteins and the chi-square based feature selection work well. A conclusion might be that the feature-selection methods work, and that it is indeed the case that some functions, or some proteins, have a higher predictive power than others. To test whether this is really the case, we compare these feature selection methods with random selection of proteins or functions. In order to evaluate the *Random* versions of the proposed methods, we did the following:

1. We chose K features randomly.
2. We used the selected method M to describe each protein in the test set based on the randomly selected features. We called the new method: M-Random.
3. We applied the naive Bayes classifier to calculate the F-measure values.
4. We repeated the steps 1 to 3 fifty times, and report the average of the F-measure values.

We assign $K = 100$ and $K = 10$ for the functional and the structural context methods, respectively. Figure 4.7 compares our proposed methods with their corresponding

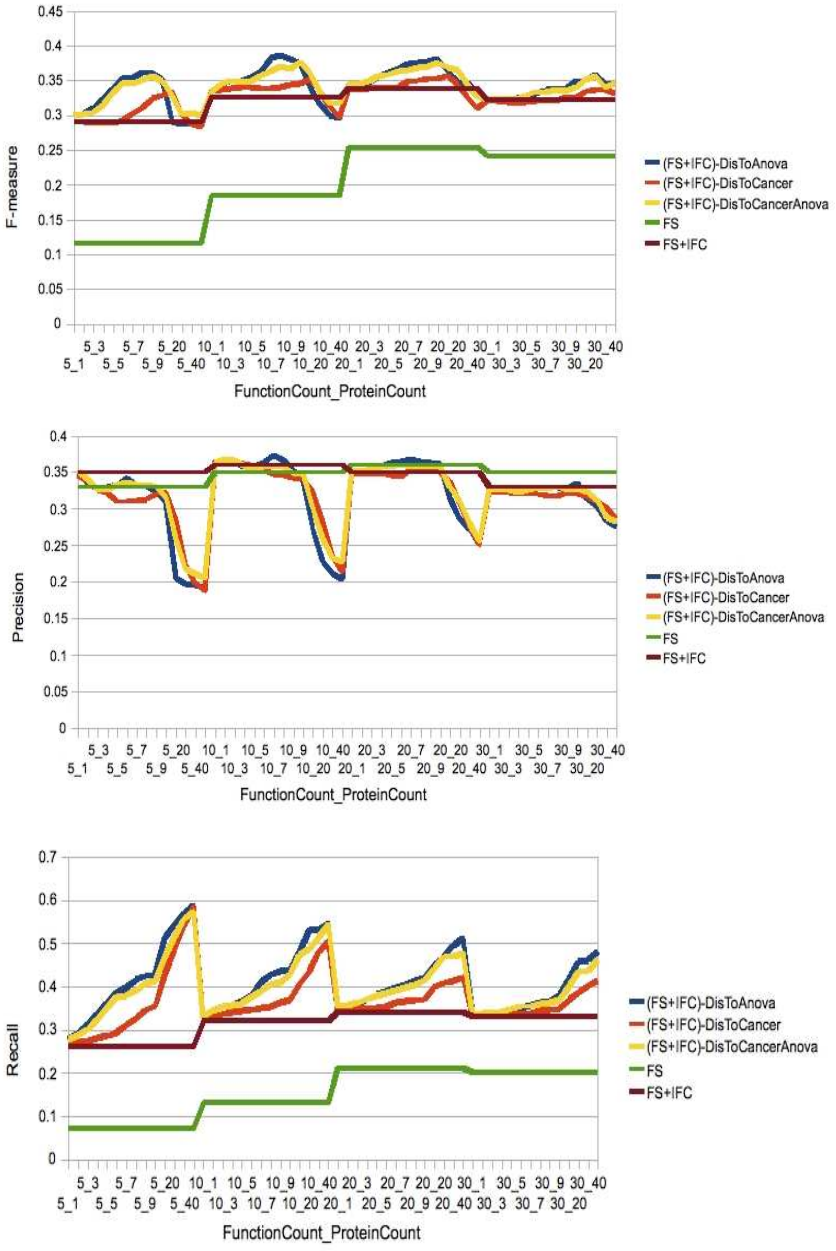

Figure 4.4: Comparing different Integrated methods. Comparing to functional and structural context methods, the integrated method gives rise to more cases (17 out of 52 in $(FS + IFC)$-DisToCancerAnova, as opposed to 0 out of 52 in $FS + IFC$) with F-measure over 35% (and up to 39% in one case).

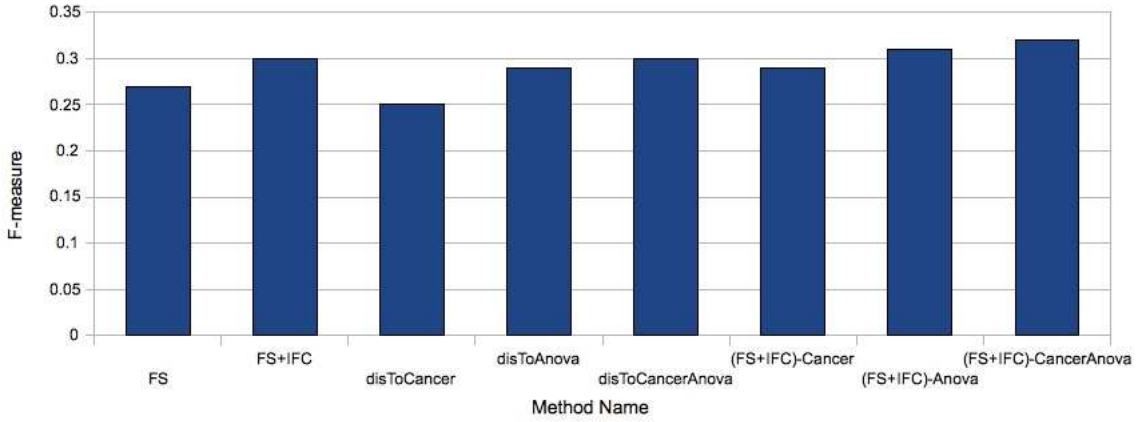

Figure 4.5: Comparison of all the proposed methods with each other. The method (FS+IFC)-DisToCancerAnova which considers network contextual information outperforms all other proposed methods.

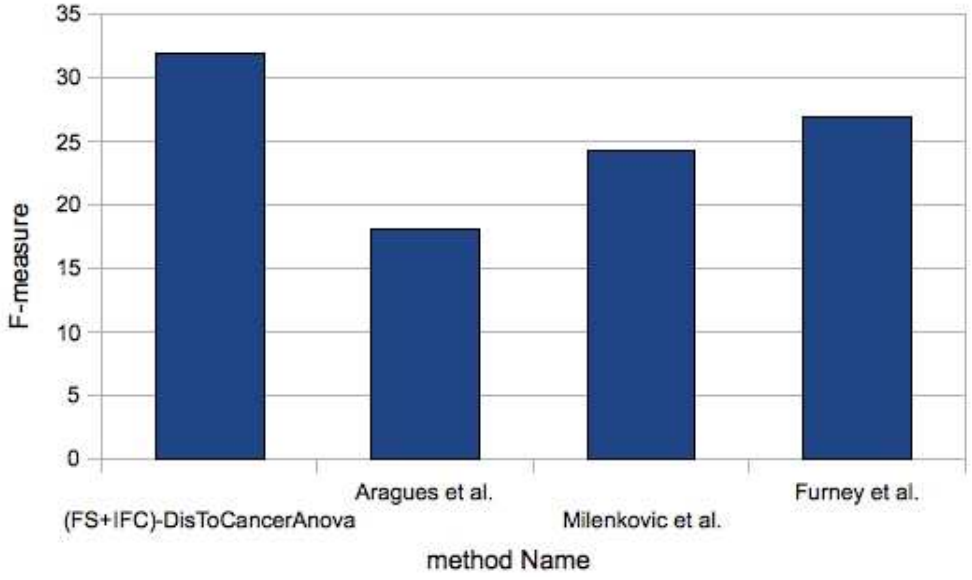

Figure 4.6: Comparing with previous methods. The method (FS+IFC)-DisToCancerAnova which considers network contextual information outperforms the previous methods (Furney et al. [29], Aragues et al. [3] and Milenkovic et al. [77]), by 5%, 13% and 8%, respectively, with respect to F-measure.

Random versions. It turns out that the feature selection algorithms outperform random selection in all cases, with F-measure improvements from 5% (for disToCancer, which also selects randomly but only among proteins known to be cancer-related) up to 26% (for FS).

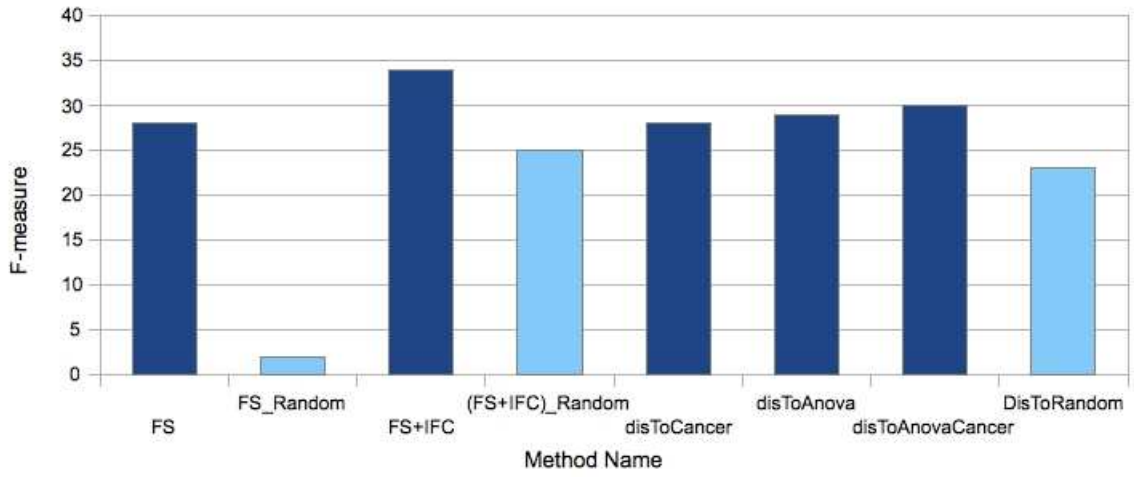

Figure 4.7: Comparing different proposed methods with their corresponding *Random* versions. The feature selection algorithm does not matter very much for the disToCancer method, with 5% improvement in F-measure over the Random version, but it matters a lot for the FS method, with 26% improvement in F-measure comparing to the Random version.

4.4.7 Capacity Identification of New Cancer-Related Proteins

The following steps were performed for predicting new cancer-related proteins:

1. A new training set was built containing all the proteins annotated as being involved in cancer (positive set) in addition to 500 randomly selected proteins (negative set).
2. A test set was built containing all the remaining proteins in the network.
3. 100 functional features were selected based on the $FS + IFC$ method.
4. 10 structural features were selected based on the ANOVA method.
5. Train set and test set were described based on the selected features.
6. The naive Bayes classifier was applied for ranking the proteins in the test set.

7. CiteXplore[68] was used to search for the high-ranked candidate proteins in the literature.

Table 4.9 lists the highly ranked newly identified cancer-related proteins. Given that, since the compilation of our dataset, novel literature linking genes to diseases (such as, in this case, cancer) have been identified, we attempted to find literature evidence for our novel gene-cancer links. As evident from Table 4.9, we found such evidence in a surprising number of at least 18 of the 20 highest-ranking genes that were not annotated in this way in the training dataset.

As can be seen, the majority of genes is now associated with breast cancer (14 out of 20), which is likely due to the fact that genotyping is currently routinely performed in this cancer type due to the different personalized treatment options available. The three genes with the least current literature information linking them to cancer are CORO2A (coronin, actin binding protein, 2A), DAZ1 (deleted in azoospermia 1) and CRSP7 (cofactor required for Sp1 transcriptional activation, subunit 7, 70kDa; now MED26, mediator complex subunit 26). However, CORO2A is involved in cell cycle progression which makes its link to cancer at least plausible. DAZ1 is involved in spermatogenesis, and it is hypothesized to bind to the 3'UTR of mRNAs to regulate their translation. While involvement of this gene in adult cancers is probably not the case, a link to regulation and cell cycle progression is also given here. Likewise, CRSP7/MED26 is a cofactor required for transcriptional activation of RNA polymerase-II dependent genes - hence, while unspecific, the link of the highest ranked genes with respect to their involvement in cancer gives a consistent link to transcriptional and, more general, cell cycle regulation events.

Overall, we were able to find literature evidence for most genes predicted to be involved in cancer, but not annotated in this manner in our training dataset. This underlines the quickly-evolving knowledge in the molecular biology field, but it also gives us more confidence that we are prospectively able to identify cancer-related genes with the approach described in the current work.

4.5 Discussion

Previous work on predicting disease-related proteins based on PPI networks has mostly focused on the functional information about the protein for which a prediction is made, or proximity of known disease-related genes in the PPI network. Several methods have been described that take into account more general features related to the graph structure, with good results. In this article, we introduce two new types of features, reflecting additional information: (1) the functions of proteins interacting with the target protein; (2) the relative position of the target protein with respect to specific other proteins, as measured by shortest-path distance. Our results indicate that:

1. Functions of proteins interacting with the target protein are informative: they offer additional information over the functions of the target protein itself. This

Table 4.9: **Capacity Identification of New Cancer-Related Proteins**

Index	Protein	Cancer Types Identified in CiteXplore	References
1	ITGAV	Breast Cancer	[11, 42]
2	CTNND2	Cervical, Prostate, Urinary Bladder	[70, 125, 46]
3	CORO2A	—	—
4	SMAD1	Breast, Colon, Lung, Prostate, Rectal, Renal cell	[62, 86, 67]
5	RPS6KB1	Breast, Colon or Rectal ovarian	[114, 90, 64]
6	VIL2	Breast and Prostate	[93, 110]
7	FST	Breast, Gastric, Lung, Prostate, Stomach, Thyroid	[85, 10, 108]
8	HSP90AA1	Gastric, Lung	[14, 115]
9	PPP2CA	Breast, Colon, Lung, Prostate	[7, 4, 128]
10	SUMO1	Breast, Lung, Prostate	[41, 80, 55]
11	SKP1A	Esophageal	[84]
12	EIF4EBP1	Breast, Colon, Head, Neck, Ovarian, Prostate	[135, 107, 5]
13	DAZ1	—	—
14	CRSP7	—	—
15	TGFB3	Breast,Colon, Prostate, Pancreatic	[112, 34, 63]
16	FHL2	Breast, Colon, Gastrointestinal, Liver, Prostate	[40, 82, 131]
17	TLN1	Breast, Prostate	[61, 94, 105]
18	GFI1B	Breast, Gastric, leukemia, Ovarian	[53, 134, 74]
19	IGFBP7	Breast, Cervical, Colorectal, leukemia, Liver, Lung, Neck, Thyroid carcinogenesis	[17, 44, 33]
20	COL4A2	Breast, Gastric, Lung, Pancreatic	[113, 8, 43]

is visible both in the expert interpretation of the results and in the predictive accuracy of the method.

2. A relatively small number of GO functions suffices to obtain maximal predictive accuracy.

3. Shortest-path distances to selected fixed proteins in the network are relevant, even more relevant than shortest-path distances to other disease-related proteins;

4. A small number of such fixed proteins (10, in our experiments) is sufficient to obtain good predictive power;

5. The χ^2 and ANOVA measures for selecting relevant functions, respectively proteins, yield interpretable results.

6. A simple and efficient machine learning method (here Naive Bayes) that uses a combination of functional information about the neighbors and shortest-path

distance to specific proteins, predicts cancer-related proteins with higher accuracy than any previous PPI-based methods.

7. What is particularly remarkable of the current work is that not only our classification results improve upon previous methods, but that also our 'false' positive predictions could in many cases be verified to be linked to cancer in more recent literature. Namely, we analyzed a list of 20 newly found cancer-related proteins that were identified by our method, and we find that virtually all of them (at least 18 out of 20) could be linked to cancer in scientific publications.

We concluded from this that the proposed features are informative for predicting cancer-related proteins as they increase the accuracy of predictive models and have a biological interpretation.

4.6 Acknowledgments

This research is funded by the Dutch Science Foundation (NWO) through VIDI grant 639.022.605. The authors thank Tijana Milenkovic for her cooperation.