# Chapter 5

# Predicting Cancer-Related Proteins using Interaction-based Features

Based on

Hossein Rahmani, Hendrik Blockeel and Andreas Bender: Interaction-based feature selection for predicting cancer-related proteins in protein-protein interaction networks. In: Proceedings Fifth International Workshop on Machine Learning in System Biology (2011)

## 5.1 Introduction

The task of predicting in a protein-protein-interaction (PPI) network which proteins are involved in certain diseases, such as cancer, has received a significant amount of attention in the literature [29, 66]. Multiple approaches haven been proposed, some based on graph algorithms, some on standard machine learning approaches. Machine learning approaches such as Milenkovic et al.[77], Furney et al. [29], Li et al. [66], Furney et al. [30] and Kar et al. [52] typically use a feature-based representation of proteins as input, and their success depends strongly on the relevance of the selected features. In earlier work it has been shown that the Gene Ontology (GO) annotations of a protein have high relevance. For instance, Li et al. [66] found predictive performance to depend only slightly on the chosen machine learning method, but strongly on the chosen features, and among many features considered, GO annotations turned out to be particularly important.

In previous work, when a protein $p$ is to be classified as disease-related or not, the GO annotations used for that prediction are usually those of $p$ itself. In this paper, we present a new type of GO-based features. These features are based not on the GO annotation ("function") of a single protein, but on pairs of functions that occur on both sides of an edge in the PPI network. We call them *interaction-based features*.

## 5.2 Interaction-based feature selection

A PPI network is a graph where nodes are proteins and an edge between two nodes indicates that those two proteins are known to interact. In our application, proteins in the training set are also labeled as cancer-related or not (supervised learning). Additionally, each protein $p$ is annotated with a vector $FS(p)$ that indicates the functions that $p$ has according to the Gene Ontology. Let $F = \{f_1, \ldots, f_{|F|}\}$ be the set of all functions in GO. $FS(p)$ is then an $|F|$-dimensional vector with $FS_i(p) = 1$ if protein $p$ has function $f_i$, and $FS_i(p) = 0$ otherwise.

Several authors [29, 66] propose to use a $\chi^2$-based feature selection method to select the most relevant GO terms. Let $C$ and $\bar{C}$ be the set of proteins that are cancer-related ($C$) or not ($\bar{C}$), and let, for each $f_i$, $P_i$ be the set of proteins annotated with $f_i$ and $\bar{P}_i$ the set of proteins not annotated with it. With $a = |C \cap P_i|$, $b = |C \cap \bar{P}_i|$, $c = |\bar{C} \cap P_i|$ and $d = |\bar{C} \cap \bar{P}_i|$, we have

$$\chi^2(f_i) = \frac{(ad - bc)^2 * (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \tag{5.1}$$

Selecting individual discriminative functions based on equation 5.1 does not consider the network topology and the way different functions interact with each other in the network. Recent approach by Rahmani et al. [99] showed that considering Collaborative Functions: Pairs of functions that frequently interface with each other in different interacting proteins, improves the prediction of proteins functions. For the task of predicting cancer-related proteins, it is not impossible that a function $f_i$

does not correlate itself with cancer-involvement, but when a protein with function $f_i$ interacts with a protein with function $f_j$, this interaction may be an indication of the former protein being involved in a cancer.

To be able to take into account the information in the interactions, we here define new features $f_{ij}$. These do not describe nodes, but directed edges between nodes. Although edges in a PPI network are undirected, we can see them as pairs of directed edges. A directed edge $p \rightarrow q$ is considered positive if $p$ is a cancer-related protein, and negative otherwise. By definition, $f_{ij}(p \rightarrow q) = 1$ if $FS_i(p) = 1$ and $FS_j(q) = 1$, and 0 otherwise. If $C$ is the set of positive edges, $\bar{C}$ the set of negative edges, and for each feature $f_{ij}$, $P_{ij}$ is the set of edges for which $f_{ij} = 1$ and $\bar{P}_{ij}$ is the set of edges for which $f_{ij} = 0$, then the $\chi^2$ value of $f_{ij}$ can be defined exactly as above (substituting $f_{ij}$ and $P_{ij}$ for $f_i$ and $P_i$ in the formulas for $a$, $b$, $c$, $d$ and $\chi^2$). Intuitively, an $f_{ij}$ with high $\chi^2$-value is relevant for the class of the protein on the $i$-side.

The $f_{ij}$ features describe edges, but we need instead features that describe proteins. Therefore, we define features $F_{ij}$ as follows: $F_{ij}(p) = \sum_q f_{ij}(p \rightarrow q)$ if $FS_i(p) = 1$, and $F_{ij}(p) = -1$ otherwise. Note that by introducing $-1$ as a separate value indicating that $FS_i(p) = 0$, each $F_{ij}$ encodes implicitly the corresponding $f_i$ feature.

In this work we compare how well cancer-involvement can be predicted from: (1) a limited number of $f_i$ features, when those features are selected according to their $\chi^2$ value as defined above, and (2) the same number of $F_{ij}$ features, when those features are selected according to the following score, which combines the overall relevance of $f_i$, $f_j$, and $f_{ij}$:

$$score(F_{ij}) = \chi^2(f_i) + \chi^2(f_j) + \chi^2(f_{ij}).$$

In the following we will call the $f_i$ individual-based features, and the $F_{ij}$ interaction-based features.

## 5.3 Results

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]. Milenkovic et al. provide details on the construction of the integrated network; some statistical information is shown in Table 5.1.

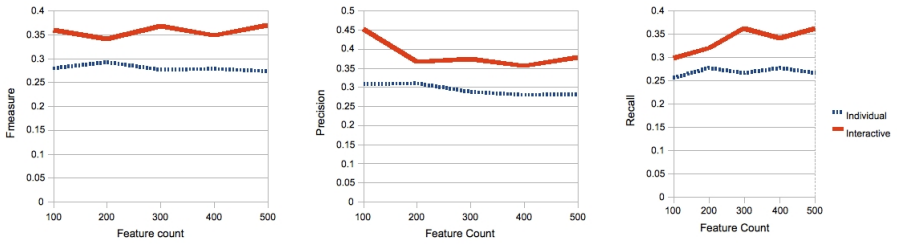
We divided the dataset into a training set containing 90%, and a test set containing the remaining 10%, of the proteins. We used information in the train set to select the $K(= 100, 200, 300, 400, 500)$ highest scoring individual-based, respectively interaction-based, features. Then, we described each protein in the test set based on the selected features and finally, we applied the Naive Bayes classifier for predicting cancer-related proteins.

Figure 5.1 compares our interaction-based features with the individual-based features with respect to the Fmeasure, Precision and Recall metrics. Our proposed method outperforms the individual-based method with 7.8%, on average, with respect to Fmeasure. This confirms our assumption about the usefulness of considering

| | |
|---|---|
| Number of proteins | 10,282 |
| Average Degree | 9.201 |
| Min Degree | 1 |
| Max Degree | 272 |
| Number of Cancer Genes | 939 |

Table 5.1: Statistical information of union of three human PPI datasets: HPRD [91], BIOGRID [116] and Radivojac et al. [97].

network interactions in feature selection. Table 5.2 lists five high-ranked function pairs; it shows that the functions in these pairs are not necessarily among the highest ranking functions with respect to their own $\chi^2$.




(a) Fmeasure (b) Precision (c) Recall

Figure 5.1: Comparing interaction-based feature selection with protein-based feature selection with respect to the Fmeasure, Precision and Recall metrics. Interaction-based feature selection outperforms the protein-based method with 7.8%, on average, with respect to Fmeasure.

What is interesting about Table 5.2 is that terms from two of the ontologies used, namely 'Molecular Function' as well as 'Biological Process', are selected using our feature selection method. This is the case both for pairs of terms from the same ontology, as well as for pairs of terms taken from both ontologies. More explicitly, GO terms 5515 and 3700 relate to 'protein amino acid binding' and 'DNA binding transcription factor activity', and are hence related to cellular replication (first entry in Table 5.2). Subsequent entries have slightly different character though, such as relating protein binding (GO term 5515) to events such as signal transduction (GO term 7165), and they are hence alerting to the particular kinds of proteins that are often involved in cancer, namely kinases (such as EGFR) involved in a large number of signaling processes in the cell. It is interesting that GO terms 60571, and also 1823 and 1656 are returned by our analysis, the former relating to 'morphogenesis of an epithelial fold', and the latter two to different stages of kidney development. Hence, some of the terms returned can also be seen as tissue-specific as well as organ-specific, and in this way a more subtle differentiation of ontology annotations can be achieved

| $f_i$ | $f_j$ | $Rank(\chi^2(f_i))$ | $Rank(\chi^2(f_j))$ | $Rank(score(f_i, f_j))$ |
|---|---|---|---|---|
| GO-0005515 | GO-0003700 | 5 | 6 | 1 |
| GO-0005515 | GO-0007165 | 5 | 46 | 2 |
| GO-0060571 | GO-0001656 | 175 | 17 | 3 |
| GO-0060571 | GO-0001823 | 175 | 105 | 4 |
| GO-0060571 | GO-0050768 | 175 | 170 | 5 |

Table 5.2: Five high-score interactive function pairs. Function members of interactive pairs are not necessarily among the functions with high chi-score value.

than by using single terms alone.

## 5.4 Conclusions

Earlier work showed that Gene Ontology annotations of a protein are relevant for predicting whether it is involved in cancer. In this work we have shown that predictive accuracy can be improved significantly by combining this information with the information contained in the topology of a PPI network. Although the combination of GO-based features and features based on network topology has been considered before, the idea of attributing GO-based features to edges, rather than nodes, is novel, and is shown here to substantially improve predictive accuracy, and to identify functional interactions for which the involved functions would not normally be found relevant by themselves.

### Acknowledgements

This research was funded by the Netherlands Organisation for Scientific Research (NWO) through a Vidi grant.