Domain Expertise–Agnostic Feature Selection for the Analysis of Breast Cancer Data

Susanna Pozzoli[a,d,*], Amira Soliman[b], Leila Bahri[a], Rui Mamede Branca[c], Sarunas Girdzijauskas[a] and Marco Brambilla[d]

[a] KTH Royal Institute of Technology, Stockholm, Sweden
[b] RISE SICS, Stockholm, Sweden
[c] Karolinska Institutet, Stockholm, Sweden
[d] Politecnico di Milano, Milan, Italy

ARTICLE INFO

Keywords:
Breast Cancer
Clustering
Clustering Performance Evaluation
Dimensionality Reduction
Feature Selection
Proteomics
Unsupervised Learning

ABSTRACT

Progress in proteomics has enabled biologists to accurately measure the amount of protein in a tumor. This work is based on a breast cancer data set, result of the proteomics analysis of a cohort of tumors carried out at Karolinska Institutet. While evidence suggests that an anomaly in the protein content is related to the cancerous nature of tumors, the proteins that could be markers of cancer types and subtypes and the underlying interactions are not completely known. This work sheds light on the potential of the application of unsupervised learning in the analysis of the aforementioned data sets, namely in the detection of distinctive proteins for the identification of the cancer subtypes, in the absence of domain expertise. In the analyzed data set, the number of samples, or tumors, is significantly lower than the number of features, or proteins; consequently, the input data can be thought of as high-dimensional data. The use of high-dimensional data has already become widespread, and a great deal of effort has been put into high-dimensional data analysis by means of feature selection, but it is still largely based on prior specialist knowledge, which in this case is not complete. There is a growing need for unsupervised feature selection, which raises the issue of how to generate promising subsets of features among all the possible combinations, as well as how to evaluate the quality of these subsets in the absence of specialist knowledge. We hereby propose a new wrapper method for the generation and evaluation of subsets of features via Spectral Clustering and modularity, respectively. We conduct experiments to test the effectiveness of the new method in the analysis of the breast cancer data, in a domain expertise–agnostic context. Furthermore, we show that we can successfully augment our method by incorporating an external source of data on known protein complexes. Our approach reveals a large number of subsets of features that are better at clustering the samples than the state-of-the-art classification in terms of modularity and shows a potential to be useful for future proteomics research.

1. Introduction

Breast cancer is the most frequent cancer type among women and one of the most common death causes worldwide. In the clinic, breast tumors are classified into five subtypes (*basal-like*, *luminal A*, *luminal B*, *HER2*, and *normal-like*) based on the status of the clinicopathological surrogates estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2, or ERBB2), and proliferation marker Ki67 (MKI67). The specific physiological features of the cancer subtypes have been found to closely follow a molecular signature comprising the expression at transcript (mRNA) level of a panel of 50 genes (PAM50) [22, 24, 27]. This classification guides the process of treatment decisions in current clinical practice. However, recent studies show that classification ambiguities still exist, and moreover suggest that the current classification system is likely incomplete, mainly because clinical routine primarily relies on immunostaining of ER, PR, HER2 and Ki67. Further still, the PAM50 gene expression signature remains insufficient to adequately stratify all tumor subtypes for the purpose of treatment [27, 15].

Cancer research has received a new dimension by the recent high-throughput mass spectrometry–based proteomic studies, where the focus is moved from transcript level gene expression to quantification at the protein level [9, 23, 26]. Accordingly, the protein quantification techniques have enabled protein-based molecular characterization of breast tumors [27, 15]. Given that the protein level is closer to the phenotype than the mRNA level, one is now able to explore protein level inter-tumor heterogeneity and thereby initiate proteomic-based classification of tumors [26, 27, 15]. These proteomic-based classifications demonstrate the importance of deep proteomic analyses, which may lead to stronger predictors of therapeutic response for better cancer treatment as well as improved patient stratification, since cellular function and pharmaceutical intervention are largely mediated at the protein level [15].

It is generally thought that all cancers have their root in mutations, abnormal changes in the DNA which lead to abnormal expression of genes and ultimately to cellular dys-

The research project has partially received funding under the Marie Skłodowska-Curie Actions (MCSA) funded project Real-time Analytics for Internet of Sports (RAIS) (Grant Agreement no. 813162). The research was also partially funded under "AI for Proteomics" initiative by RISE Research Institutes of Sweden.

*Corresponding Author
spozzoli@kth.se (S. Pozzoli)
ORCID(s): 0000-0002-6899-6209 (S. Pozzoli); 0000-0002-8753-2434 (M. Brambilla)

function [17]. Tumors can either be inherently resistant or acquire resistance to treatment, these traits also being the result of a faulty gene expression at the cellular level. Whilst the molecular signature provided by the PAM50 gene expression profile at transcript level is extremely useful to understand the biology of the breast cancer subtypes, there is often a substantial correlation gap between the transcript and its protein product, due to e.g. translational regulation or targeted protein degradation [15]. Proteomic data analysis, being a more direct assessment of the main functional molecules in a cell (i.e. the proteins) can increase the understanding of how DNA mutations can lead to malicious cell behaviour, and may help elucidate cancer mechanisms more amenable to therapeutics. This is why one should go beyond mRNA-based PAM50 classification, and look for similarity patterns among breast cancer patients at the proteomics level, which may uncover more accurate patient stratifications, eventually leading to improved treatment decisions.

Proteomic-based profiles contain a high number of protein products, usually reaching a quantitative depth of $10,000+$ proteins [15, 26]. This high dimensionality makes the tumor classification task very challenging and computationally expensive, specifically when it is required to select the most important proteins for effective disease classification. Although the field of data mining has different analytical pipelines for big data, the main concern with proteomics is data sparsity, such that the number of data samples (i.e., patients) is too small compared to the number of measured proteins. This can be described as high-dimensional small-sample problem, where we have the number of variables or features usually being much higher than the number of samples. Thus, available data mining and machine learning models of proteomic data mainly depend on supervised machine learning techniques [26, 27, 12]. For example, Tyanova et al. [26] trained three different supervised classification and feature selection models, each one separating a single breast cancer subtype from the other two subtypes included in their data set.

The problems with high dimensionality result from the fact that the set of data points becomes increasingly sparse as the dimensionality increases, which in turn makes deeper analysis of data unfeasible or unattainable. Particularly in clustering purposes, the high dimensionality affects the distance or similarity metric used to cluster data points. Meaningful clustering requires that the objects within clusters are, in general, closer to each other than to objects in other clusters. Previous work, which analyzed the behavior of distances for high dimensional data, showed that the distances between points become relatively uniform in high dimensional spaces, to the point that data clustering becomes meaningless as the notion of the closest and farthest neighbor of a point ceases to exist [2]. However, most data features are highly redundant and can be efficiently scaled down to a much smaller number of variables without a significant loss of information. This process of dimensionality reduction can be performed in two different ways: by only keeping the most relevant features from the original ones (this technique is called *feature selection*) or by finding a smaller set of new variables, each being a combination of the input features, containing basically the same information as the input variables (this technique is called *feature transformation*) [6].

Feature selection can be used to transform high dimensional proteomic data into a reduced representation. However, it is a challenging task to find a reduced meaningful representation that maintains the intrinsic properties of the data, such that the observed properties and underlying patterns become more pronounced using the reduced representation. In this study, we address the challenging question of identifying relationships between cancer cell behavior and specific groups of proteins. Differently from existing work that focuses on analyzing the performance of using quantitative protein levels to cluster the tumors and the accuracy of this clustering compared to the current consensus of breast cancer subtypes, our objective in this study is to propose unsupervised methods to extract groups of proteins that can be linked to idiosyncrasies of the disease without the need of prior expertise knowledge.

Due to the domain of the analysis, it is important for the results and thus for the methodology to be interpretable, because we cannot take it for granted that certain proteins are connected to breast cancer. In other words, we must be able to retrace our steps, to explain why we select the one protein instead of the other. As Hutson writes in [14], Artificial Intelligence (AI) has received a keen interest in recent years. It has expanded the frontiers of computer science, but at the same time its marvellous results have led people to employ it as a passepartout that is arbitrarily used to execute tasks for which AI is not really necessary. This is why, our wrapper method is constructed on top of classical machine learning.

Using 45 tumor samples [15] from the Oslo2 landscape Breast Cancer cohort we perform our domain expertise–agnostic analysis by creating a similarity graph of proteins. The basic idea is to construct a graph from the proteomic profiles where each vertex represents a protein, and each weighted edge represents the similarity between two proteins with respect to their measured levels in the tumor samples. Then, this graph is fed into a graph clustering algorithm that partitions the graph into clusters, such that proteins that belong to a group should be similar (or related) to one another and different from (or unrelated to) the proteins in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better the clustering. We propose two different methodologies for creating this similarity network among the proteins, the first methodology is based on the quantified protein levels, whereas the second one is created using protein complexes, specifically the CORUM database [11].

Our proposed unsupervised feature selection pipeline is composed of three steps as shown in Figure 1. We start with the *feature filtering* step by selecting the set of proteins having the highest variance across the 45 patients. We only want to consider the proteins that show different values with different patients (i.e., present a pattern among the patients), while ignoring proteins that carry no signal and show al-

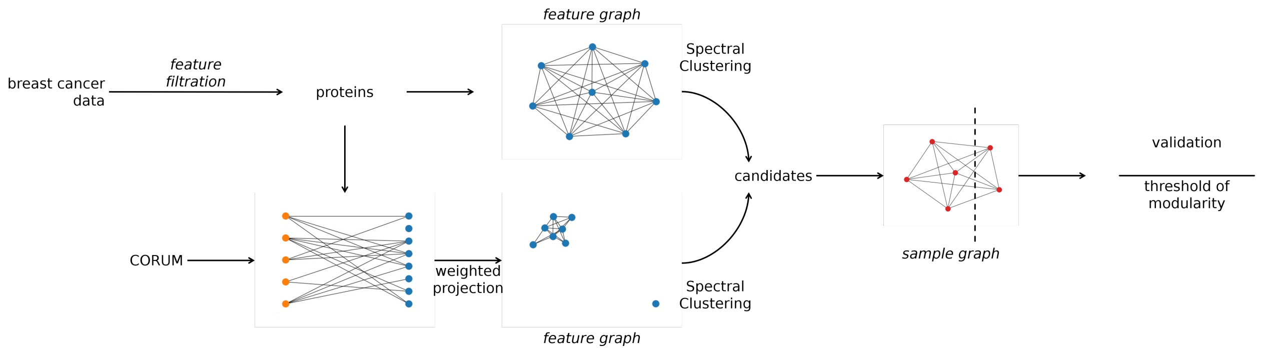

Figure 1: The proposed unsupervised feature selection pipeline for breast cancer proteomic data analysis. First we perform *Feature Filtering* step by focusing the analysis on most varying proteins. Afterwards, we create two similarity graphs to proceed further with *Candidate Generation* step. Our last step is *Candidate Evaluation* that is based on calculating the modularity scores.

most same value across the set of patients. Therefore, we calculate the *z*-score for all the proteins. Specifically, we calculate the population mean as the mean of all entries in our data set, then for each protein we calculate the average number of standard deviations its associated values with different patients are away from the population mean. Having the *z*-score being calculated for each protein, we rank the proteins and select the set of most varying proteins and use them to create the similarity graph.

Having created the similarity graph, we start our second step, *i.e.*, *candidate generation*, by performing the process of graph clustering to split the input graph into multiple groups. The graph clustering step can produce too many possible ways of grouping the proteins, thus we compute modularity score for the extracted subsets of proteins [3]. We use the modularity score as measure to indicate the strength of splitting the tumors into two groups using every extracted protein group. Afterwards, we pick the protein groups with the highest modularity and we define this process as *candidate evaluation* in our analysis pipeline. These candidates represent the groups of highly correlated proteins that can be used as a reduced dimensional representation of the data to perform further analysis to uncover more patient stratifications. Table 1 lists some of the candidate protein groups that have the highest modularity score.

Our results show that the protein groups with highest modularity are associated with biological phenotypes distinguishing the different tumor samples. Namely, one of the most notable clusters contains proteins that are expressed at higher levels during mitosis, underpinning cell proliferation, indicating that the tumors can be well classified in to high or low proliferation. And whereas this property, the proliferation state, is a rather well known classifier in breast and other cancers, our current analysis also reveals more subtle clusters associated with other biological processes, such as extracellular matrix homeostasis, lipid metabolism, and immune response.

Accordingly, we can describe the contributions of this paper as follows:

- protein similarity network that can be created using different data sources, e.g., proteomic data as well as protein complexes database;
- clustering pipeline to reduce the dimensionality of the similarity graph, and selecting groups of proteins that capture patterns across cancer cells; and
- biological interpretation of the top-scoring proteins, in light of the differences between the samples, which highlights potential benefits of proteome data analysis.

Outline. The following will explain the disposition of the article. Section 2 provides background on breast cancer subtypes as well as a description of the proteomics data set used in this work. Section 3 discusses related work; whereas, Section 4 presents the research methodology used. Section 5 presents the results, which Section 6 discusses from a biological perspective. Section 7 presents the future work and Section 8 concludes the article.

2. Breast Cancer and Proteomics Data

In this section we provide a brief overview on breast cancer molecular subtypes as known in the literature. We also describe the proteomics data set retrieved from Karolinska Institutet and analyze its content based on these known subtypes.

2.1. Breast Cancer Subtypes

Researchers have been studying the classification of breast cancer based on molecular characteristics, and how this could be useful in planning treatment and developing new therapies [15]. The complex profile of each subtype is determined using both molecular information and genetic expression profiling from tumor cells. Although there are references about some less common molecular subtypes, such as *Claudin-low* [7], most recent studies agree on dividing breast cancer into five major molecular subtypes: *basal-like*, *luminal A*, *luminal B*, *HER2*, and *normal-like*. These subtypes are mostly used in research settings with the aim of guiding better and more personalized treatments.

Recent developments in proteomics have allowed researchers and practitioners to quantify proteins in tumor cells with unprecedented success. In [15], researchers have been able, for the first time, to identify the molecular subtype of tumors based on analyzing breast tumor proteomes, where 9, 995 proteins have been quantified across all tumors. Another outstanding outcome of that proteomes analysis was the ability to further subdivide *basal-like* and *luminal B* tumors. This result suggests that the currently adopted molecular classification of breast cancer tumors can be further expanded to find even more fine-grained subtypes which could guide more personalized and better targeted treatments. Towards achieving this, this work aims to perform a more thorough analysis of the data, based on state-of-the-art machine learning mechanisms.

2.2. Breast Cancer Proteomics Data

The data set covers a panel of 45 breast cancer samples consisting of 9 samples per each cancer subtype (i.e. *basal-like*, *luminal B*, *HER2*, *luminal B*, and *normal-like*). Each sample has a proteomic profile of 9, 995 features, each representing the quantification of a specific protein, where the possible values range from 0.00478 to 30.087438. Numbers are not absolute values but are instead ratios, as determined by dividing the abundance of each protein in a given sample by the average of the forty-five breast cancer tumors, and as a result their arithmetic mean is equal to 1.0. See Johansson et al. [15] for details of the laboratory analyses.

There is a large number of outliers. Row-wise, standard deviations range between 0.067 and 4.735, despite the fact that the overall standard deviation is equal to 0.412, which means that we found a wide variance in some of the features. Overall, the Interquartile Range (IQR), *i.e.*, the difference between the upper quartile and the lower quartile, is $1.147 - 0.854 = 0.293$. According to Tukey [25], outliers are, by definition, data points below the lower quartile or above the upper quartile by a margin of at least 1.5 times the IQR. Here, 28, 681 values out of $45 \cdot 9,995 = 449,775$ fall within the definition of ’outlier’, in support of our working hypothesis that an anomaly in the data could indicate cancer.

Large numbers of values are concentrated in the area of the arithmetic mean and the distribution has a right skew. In order to normalize this data, we adopt the common practice of log-normalization by calculating the binary logarithm of the values. In this way, we can reduce skewness and properly weigh outliers on both sides of the distribution.

In order to better understand the relationships between the proteomics data in hand and the known sub-types of the covered samples, we provide a visualization of the state-of-the-art classification of breast cancer cells in Figure 2. Figure 2 shows the binary logarithm of the protein content, where samples are shown on the horizontal axis, divided into the five breast cancer subtypes. As can be seen on the figure, different groups of proteins highlight the presence of different breast cancer subtypes. For example, blue in the top left-hand corner of Figure 2 puts a lot of emphasis on the breast cancer cells of type *basal-like*. Also, in case of type *HER2*, there is an excess of ERBB2 and GRB7.

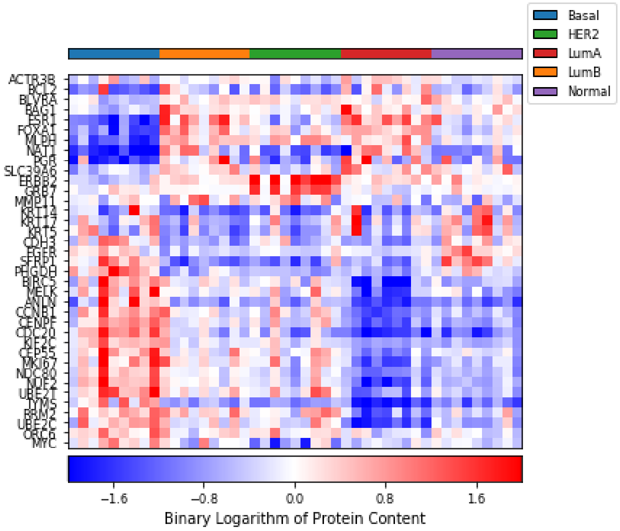

Figure 2: Representation of the five subtypes of breast cancer cells, *basal-like* (blue), *luminal B* (orange), *HER2* (green), *luminal A* (red), and *normal-like* (purple), as available from the proteomics data of the labeled samples.

3. Related Work

We review the related work from two perspectives. First, we look unto prior works on proteomic data analysis, with a focus on the ones adopting computer science approaches. We thereafter review existing works on dimensionality reduction, as our proposal for proteomic data analysis relies mostly on this area of computer science research.

3.1. Proteomic data analysis

Prior to advances in proteomic data generation, work has previously been done in the field of DNA microarray data analysis. DNA microarray measures the expression of the genes but it is a rough estimate of the actual protein content. Support Vector Machines (SVMs) are popular tool for the analysis of this kind of data as detailed in [13]. However, the focus has mainly been on the problem of recognition of cancer types, instead of subtypes, based on the gene expression. Peng et al. [21] proposes a method based on a Memetic Algorithm (MA) for the selection of the genes characteristics of each of the cancer types, such as lung cancer and leukemia. The proposed solution is a Genetic Algorithm (GA), which makes use of the crossover operator for recombination, enriched with an iterated local search. Similarly, Duval et al. [8] use a combination of a GA and a SVM. All these works have been exploiting data analysis techniques on DNA microarray data, which provided only rough approximates of protein content in tumor cells.

As the field of proteomics developed and scientists became able to generate measurements of actual protein contents in cells, many data-driven classification approaches have been proposed using quantitative proteomics to examine the mapping of clustering data using some of the measured proteins to the well-established breast cancer subtypes known using mRNA-based PAM50 markers. Works such as Tyanova et al. [26] and Johansson et al. [15] provide different types of analysis performed at the protein level in order to come up with classification frameworks that would map the samples in the proteomics data to the known cancer molecular subtypes. For instance, in Tyanova et al. [26] the authors propose cancer subtype classification framework using a data set consisting of forty breast cancer samples having proteomic profiles with depth 7,000+ proteins in each sample. Their proposed framework incorporates several supervised machine learning methods to perform the tasks of classification, feature selection and cross-validation. The authors employ SVM for classification purposes, such that they train three different classifiers, each one separating one breast cancer subtype from the other two. In the most recent work in Johansson et al. [15], the authors employ hierarchical clustering on breast cander proteomic data to re-identify the known 5 molecular subtypes. All these works provide evidence that cancer subtyping can be performed based on analysing proteomics data, and may provide more fine grained classes than pure molecular or genetic based analysis in a laboratory. However, almost all these works employ supervised, or at best semi-supervised, machine learning mechanisms, which limits the search space to already known knowledge about cancer subtypes and their molecular causes. The novelty of our performed work is the employment of domain knowledge-agnostic dimensionality reduction with purely unsupervised feature selection for the identification of groups of tumors that share common proteomics-level characteristics, without being biased by prior knowledge on identified molecular subtypes.

3.2. Dimensionality Reduction

In general, *dimensionality reduction* is the task of moving from an m_1-dimensional space to an m_2-dimensional space, where $m_2 < m_1$. Of course, the low-dimensional space is fully expected to represent the high-dimensional space. There are two categories of methods. According to Cunningham [6], while *feature transformation* creates brand-new dimensions, starting with or without the existing dimensions, *feature selection* is performed on the existing dimensions without affecting the interpretation of the inference.

3.2.1. Feature Selection

In general, *feature selection* is the task of selecting m_2 dimensions out of m_1. There are three categories of methods: *filter*, *wrapper*, and *embedded* methods [16]. Filter methods are being used to discard the irrelevant features right away. To do so, firstly, the features are ordered according to numbers, such as the information gain and the Pearson's correlation coefficient, and secondly, the features with significance below a given threshold are ignored. After neglecting the insignificant feature, any statistical learning model would be fitted only once using only the selected features. On the other hand, Wrapper methods are being used to limit the search space. Heuristics, such as stepwise selection and GAs, are used for candidate generation. The whole idea of the algorithm is to find the most promising candidates. As for embedded methods, they try to train the statistical learning model and to select the features simultaneously.

The same distinction holds both for the supervised and the unsupervised setting. The majority of previous work in the field of feature selection focuses on the supervised scenario [4], because in case of unsupervised methods, such as clustering, it is difficult to define the criteria according to which the feature selection process is considered successful. As a matter of fact in case of classifiers it is possible to measure the accuracy achieved when considering a given subsets of features but this is not possible in case of clustering where the true labels of the points are unknown. Some alternatives to measure the quality of clusters could be deployed, but hugely depend on the specific domain and settings under hand. In our work, we deploy unsupervised feature selection using Spectral Clustering, with modularity as the criteria based on which cluster quality is evaluated.

4. Methodology and Materials

We propose a wrapper method of unsupervised feature selection for the generation and evaluation of candidate subsets via Spectral Clustering and modularity,respectively. We define candidates to be feature subsets, in this case protein subsets, with good inter-connectivity according to a given similarity network. This wrapper method is composed of three steps, as follows:

- *Feature Filtering*. In order to reduce computational costs of later steps, feature filtering is used. In this work we consider features whose z-score, which is a measure of a distribution's spread, is high, as they are more likely to highlight distinct behaviour of a subset of tumor samples. Therefore, this first step consists of selecting proteins with z-score higher than a specified threshold. This threshold could be identified based on the data in hand. It is also worth mentioning that this step is optional, it is meant for efficiency considerations only, and it may be skipped if it is difficult to identify a reasonable threshold for this z-score based filtering.

- *Candidate Generation*. So as to generate candidates, in the absence of prior knowledge, we resort to unsupervised learning. In general, clustering enables the detection of elements that form groups that are closer to each other compared to the rest of the elements in the input space. We make use of one of the state-of-the-art-clustering techniques, i.e., Spectral Clustering. Since the algorithm takes an affinity matrix as input, we convert the breast cancer data into a graph, in which the weight of the edges between the nodes, that is the features, is proportional to how affine the features are across the samples. Also, we develop a second version of this step that, after importing a repository of protein complexes, encodes the information about the collaboration between the proteins into the weighted edges in the feature graph.

- *Candidate Evaluation.* In this step, the generated feature candidates are evaluated. This is done by using each candidate as the basis for tumors classification and comparing the achieved results to those achieved using the state-of-the art breast cancer classification following the PAM50-based system. The comparison between the achieved classifications is based on the *modularity* score of the bisection of the samples.

Working hypothesis is that breast cancer subtypes are recognized by candidates, that is feature subsets, that highlight the presence of dissimilarities between samples. In other words, we assume that differences between the samples that might indicate the existence of a breast cancer subtype are easier to recognize when considering only a small fraction of the available features.

Figure 1 shows the steps of our method. *Candidate Generation* has two branches. While the top branch exploits the breast cancer data only, the bottom branch imports a small protein-to-protein interaction data set from CORUM [11]. The purpose of the second branch is to fine tune the distances between proteins also based on their known biological interactions in normal cells. ‘

4.1. Feature Filtering

Unfortunately, the generation and evaluation of candidates is a problem of great computational complexity, in terms of both time and space. As a matter of fact, for the exhaustive search, it is necessary to generate and evaluate $2^{9,995}$ different sets of candidate features. Of course, a smaller number of features will soften the requirements of the algorithm.

Taking inspiration from filter methods, we assert that it is preferable to discard the features whose information content is not significant in order that anomalies may cut through the noise. Since *feature filtering* is optional, we affirm that our method is still in the category of wrapper methods.

Proteins are listed in descending order of explained variance. Intuitively, the top n features are more likely to highlight the presence of an anomaly in the data. Therefore, it is logical to ignore all the features with low variance, because they exhibit similar measurements across all samples.

We have to sort the features by a measure of variance in order to filter in an objective way. Among other metrics, we can use the z-score, also known as standard score, as a measure of how spread is a distribution. The z-score of X is

$$z = \frac{|X - \mu_X|}{\sigma_X}, \tag{1}$$

where μ_X is the arithmetic arithmetic mean and σ_X is the standard deviation of X. Both μ_X and σ_X have been calculated on all of the $45 \cdot 9,995$ values of protein content. Basically, the z-score is equal to the difference between X and the arithmetic mean expressed in number of steps when the step is equal to the standard deviation. We added the absolute value to make the numbers comparable regardless of their being on the left or on the right of the arithmetic mean.

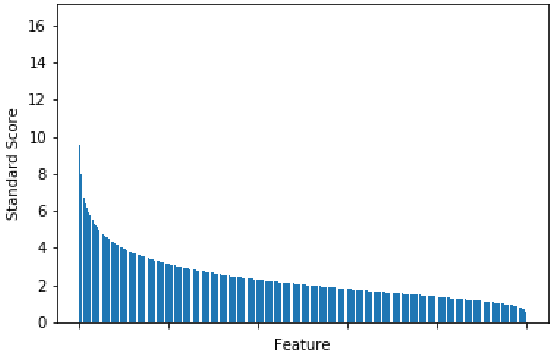

Figure 3: Vertical axis shows the standard score of the features, that is the proteins, in descending order.

The standard score of a feature is equivalent to the maximum standard score computed over all samples. Figure 3 shows the maximum absolute standard score of the features.

Features were ordered by the z-score to assess how they would explain the variance in the data. Now, it is possible to filter them. With breast cancer data, we made a choice of a threshold that is unavoidably arbitrary. For example, we could select the top n features or the features whose z-score is greater than $\mu + \sigma$, where μ is the arithmetic mean of the z-score and σ is the standard deviation of the z-score, as it was the case with the experiments whose results are detailed below.

4.2. Candidate Generation

As aforementioned, we adopt two different techniques to generate the similarity graph representing the relationships among the proteins.

4.2.1. Proteomic Data

The first *candidate generation* option generates the list of candidates by using the breast cancer data only, without external sources of information. Starting with the hypothesis that different cancer subtypes will exhibit characteristic protein signatures (which reflect the biology of the particular tumor, as well as provide hints into possible druggable targets), we expect to generate clusters of proteins whose content is different from sample to sample. Basically, it is a matter of clustering the features, instead of the samples, by transposing, which resolves the problem of our data being high-dimensional.

For clustering, our choice falls on the state of the art, i.e., Spectral Clustering [20]. However, the algorithm requires a number of clusters and an adjacency matrix, that is, a graph, as input. We have decided for a variable number of clusters, i.e., $[3, 4, \dots, 15]$, which means that, first, we divide features in three, second, we divide features in four, and so on. In the end, we compute the set of candidates as the union of the clusters of features just generated. On the plus side, it enabled us to generate overlapping clusters. However, we can expect the size of candidates to be large, especially in the cases of small numbers of clusters, such as 3 and 4. From a purely biological point of view, it is necessary to double-check our list of candidates to validate the experiment's results, because it is not certain that there will be a correspondence between breast cancer and anomalies. Of course, differences between samples might be due to conditions that are not related to cancer and there is no way we can derive it from raw cancer data. Consequently, it is imperative that the size of candidates be small. By recursion, we continue clustering the candidates whose size is over the maximum size that is possible to double-checked comfortably.

Next, it is necessary to convert the breast cancer data from a table to a graph, i.e., F. As from now, F is generally referred to as the *feature graph*. There are $9,995$ nodes in F, one per protein. It is complete, because we need to have a feature compared to all others. Weight of the edge between u and v is the affinity between the two nodes. We want the affinity between the features whose signals are either in phase or in antiphase to be maximum. For example, there are close affinities between the i^{th} feature and the j^{th} feature if the amounts both of the i^{th} protein and of the j^{th} protein are small or large at the same time as well as if the amount of the i^{th} protein is small (resp., large) and the amount of the j^{th} protein is large (resp., small). Consequently, proteins with like or opposite behaviour are expected to be in the same cluster of features. Weight of the edge between u and v is the absolute value of the cosine of the angle between u and v.

$$A_{ij} = \left| 1 - \left(1 - \frac{f_i \cdot f_j}{\|f_i\|_2 \|f_j\|_2} \right) \right| = \left| \frac{f_i \cdot f_j}{\|f_i\|_2 \|f_j\|_2} \right|, \quad (2)$$

where f_i is the value of the i^{th} feature and f_j is the value of the j^{th} feature. The possible values range from 0.0 and 1.0. Because of the absolute value, A_{ij} is 1.0 if the signals are either in phase or in antiphase – in other words, if the angle between the two features is 0° or 180°.

4.2.2. *Protein Complexes*

When generating candidates however, it is important not to lose sight of the big picture. It is a fact that proteins are not isolated in the cell. On the contrary, they collaborate on molecular functions as well as biological processes. Proteins communicate a whole range of messages, which control the workings of the cell.

There is a lot of data on the subject of Protein-Protein Interaction (PPI), but in this case we import data from CORUM [11], which is a collection of protein complexes including, but not limited to, human protein complexes. In a nutshell, a protein complex is a complex system of two or more proteins that associate physically to perform a variety of tasks. Thus, CORUM [11] allows us to augment our search space. Naturally, it is possible to add other PPI networks, as long as no domain expertise is required to join the data sets and the data structure is known.

Once again, we create a graph, but in this case, it is used to encode the participation of proteins in the protein complexes. Like before, Spectral Clustering is used to divide nodes, that is features, into candidates.

We create a bipartite graph B where the vertices are $9,995$ proteins on one side and $2,916$ protein complexes on the other side. A protein is connected to a protein complex if it is part of that protein complex. Since B is bipartite, the only edges in the graph are between proteins and protein complexes.

Candidates have no protein complexes; therefore we need to encode the further information added by the protein complexes in the weighted edges in the so-called *feature graph*. We generate a new graph, i.e., F', which is the weighted projection of B onto the proteins. There are $9,995$ nodes in F', one per protein. Two proteins u and v in F' are connected with an edge if they share at least one common protein complex in B. Weight is directly proportional to the number of shared protein complexes in B, but in this particular case it is important to penalize those protein complexes common to many proteins because the signal of high-degree protein complexes is not as strong as the signal of low-degree protein complexes. So, we use Newman's weighted projection of B onto the proteins [18]. Weight of the edge between u and v is

$$w_{u,v} = \sum_k \frac{\delta_u^k \delta_v^k}{d_k - 1}, \quad (3)$$

where d_k is the degree of node k and δ_u^k is 1 if the edge between u and k is in B. Of course, we skip the protein complexes whose degree is less than 2. For example, let us assume that there are two proteins, i.e., u and v, in B and only one protein complex, i.e., k. It is not possible that u and v share a common node k if the degree of node k is 1 – in other words, if the number of neighbors of node k is equal to 1.

F' is not guaranteed to be connected; consequently, it is possible that we will deal with distinct connected components. In a graph, a connected component is a subgraph such that there is a path between all nodes in it. We consider these connected components to be candidates, since they are, by definition, subsets of features, and we insert them in the set of candidates, too.

Here, there was a number of connected components whose size was too large to be validated manually. Therefore, we divided them into smaller candidates by clustering recursively, as detailed above.

4.3. Candidate Evaluation

At this point, we need to have our list of candidates tested in order that candidates may be put in a order. Again, assuming the principle that an abnormal amount of a given protein or proteins underlays the cancer phenotype in a given tumor, our aim is to separate the cancer samples having an above-average or below-average protein content from those having an average protein content. In other words, we want to divide the samples into two groups. On the one side, there should be samples with higher protein content; on the other side, there should be samples with lower protein content. We prefer clustering to classification, disregarding the PAM50-based breast cancer subtypes. First, there is some evidence of incompleteness of the state-of-the-art classification of breast cancer patients, and second, it may be possible that we create a cluster of samples that is incompatible with existing categories of breast cancer subtypes.

Unlike *candidate generation*, we cluster the largest eigenvectors of the affinity matrix, instead of the ones of the normalized Laplacian matrix when running Spectral Clustering because we want to maximize the average degree.

Once again, the number of clusters is an input parameter of the algorithm. Since we want to assess how good the features in the candidate under examination are at identifying clusters of samples, we set 2 as the number of clusters. The reason for this number is two-fold. First, we do not want to force a further division of a homogeneous cluster, and second, we can always increase the number of clusters. Also, it is easy to see Figure 2 as a stack of five bisections. Next, we create a complete graph, i.e., S, which is called the *sample graph*, in this case. There are 45 nodes in S, one per sample. As we have already mentioned, the weight of the edges in the graph is proportional to the affinity between the nodes in the sample graph. Our goal is to minimize the intra-cluster distance, but at the same time, to maximize the inter-cluster distance, and as a result similar samples have an affinity, unlike opposite samples. The weight of the edge between u and v is calculated on the value of the cosine of the angle between u and v, plus 1.0.

$$A_{ij} = 2 - \left(1 - \frac{s_i \cdot s_j}{\|s_i\|_2 \|s_j\|_2}\right) = 1 + \frac{s_i \cdot s_j}{\|s_i\|_2 \|s_j\|_2}, \quad (4)$$

where s_i is the value of the i^{th} sample and s_j is the value of the j^{th} sample. The possible values range from 0.0 to 2.0.

We cluster the sample graph generated according to the features in the candidate under evaluation, one candidate at a time. It is necessary to compare the quality of clusters numerically. There is a variety of clustering performance metrics, but in this particular case our choice fell on the modularity score. Fortunato [10] described modularity as a thorough performance metric, because it compares the community structure of a graph with the community structure of a random graph, numerically. According to Blondel et al. [3], the modularity score is

$$Q = \frac{1}{2m} \sum_i \sum_j \left(A_{ij} - \frac{k_i \cdot k_j}{2m}\right) \delta\left(c_i, c_j\right), \quad (5)$$

where m is the sum of the weight of the edges in B, A_{ij} is the weight of the edge between i and j, k_i is the weighted degree of node i, and c_i is the label of node i. δ is the Kronecker delta. The possible values range from -1.0 to 1.0. It is worth remembering that the modularity score of a random graph is zero. Equation 6 is Equation 5 in matrix form.

$$Q = \frac{1}{2m} \operatorname{tr}\left(\mathbf{s}^T \mathbf{B} \mathbf{s}\right), \quad (6)$$

where

$$B_{ij} = A_{ij} - \frac{k_i \cdot k_j}{2m} \quad (7)$$

and s_{ij} is 1 if the predicted label of the i^{th} sample is j. Q is, by definition, the trace of $\mathbf{s}^T\mathbf{B}\mathbf{s}$ divided by $2m$, but in this particular case we are interested in the best cluster of samples and, consequently, we return the maximum item on the principal diagonal.

According to Newman [19], **B** is a zero-sum matrix, column-wise and row-wise, and as a result in case of bisection, modularity is not sufficient to compare the quality of clusters. So, there need to be at least two clustering performance metrics to assert that the one cluster is better than the other cluster. For example, it is possible to measure the internal and/or external connectivity with metrics such as *average degree* and *conductance*.

Candidates are sorted by modularity; however, the translation of the list of fifty transcripts into the list of thirty-seven proteins found in the current proteomics data set [15] enabled us to set the threshold of modularity so that candidates may be benchmarked against the state-of-the-art classification of breast cancer tumors. It is not easy to make a fair comparison to the PAM50-based classification of breast cancer cells, because the number of clusters is different. Firstly, we thought about setting the threshold of modularity to the modularity score of the bisection of the sample graph generated according to the state-of-the-art classification of breast cancer cells. However, it was surprisingly low. We could increase the number of clusters from 2 to 5, but bisection is perfectly compatible with our working hypothesis that it is possible to recognize breast cancer subtypes because they are subsets of samples whose amount of certain proteins is significantly higher or lower than average. We tried to compute the modularity score of the subsets of the thirty-seven proteins that match the known breast cancer subtypes, visually. However, it was hard to tell which proteins are mainly responsible for which breast cancer subtypes, apart from *basal-like* and *HER2*, to some extent, and as a result we reverted to our first choice of modularity threshold.

5. Experimental Results

Experiments were conducted both with and without *feature filtering*, but we report the result of the experiment to test all three steps, because *feature filtering* sped up our methods, but at the same time it did not have repercussions for the quality of the result.

Figure 4 shows the result of the experiment. Note that the orange solid line represents the adopted threshold of modularity, *i.e.*, 0.03, which is equal to the modularity score of the thirty-seven proteins on which the state-of-the-art classification is based. Here, the maximum size of the candidate is equal to 42, which is still large, but we noticed that after a while, candidates started to repeat themselves and thus we stopped clustering.

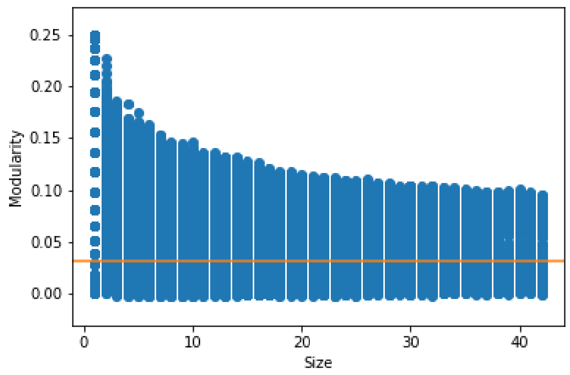

Figure 4: Modularity is on the vertical axis, and the size of the candidate is shown on the horizontal axis.

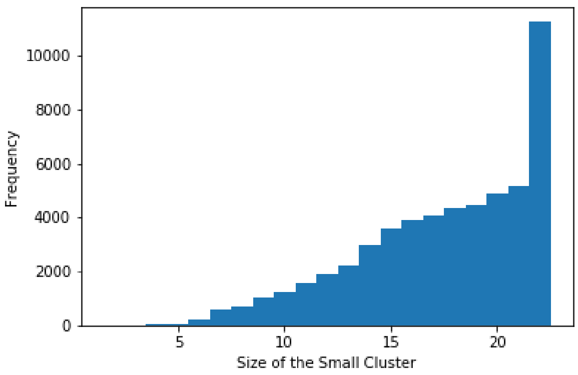

Figure 5: Vertical axis shows the frequency of the size of the small cluster.

As detailed above, we evaluated different threshold options, because in spite of the fact that the PAM50-based classification of breast cancer cells has proven to be a valuable tool in the clinic, the modularity score it achieves is surprisingly low. We suspect it is due to the larger number of candidates, *i.e.*, 5 instead of 2, but at the same time the candidates are evaluated on the bisection of the samples for the reasons stated above. Therefore, we proceeded to compute the modularity score of the subsets of the thirty-seven proteins that, visually, match the breast cancer subtypes and, actually, the modularity score was higher. For example, in case of *HER2*, the modularity score of the bisection of the sample graph generated according to ERBB2 and GRB7 is equal to 0.11, which is 3.5 times better than the score of all thirty-seven proteins. However, it is not easy to map subsets of proteins to known breast cancer subtypes, although there are only 37 proteins. For this reason, we opted to set the threshold to the modularity score of all thirty-seven proteins.

Starting with 9,995 proteins, the number of generated candidates was over 700,000, in spite of Branch and Bound (BB), which was adopted to mark the boundary of the search space. With *feature filtering*, 1,332 features out of 9,995 were selected. We keep features whose z-score is over $\mu+\sigma$, where μ is the arithmetic mean and σ is the standard deviation of the z-score. Candidates decreased in number, *i.e.*, 54,447, but at the same time there were not any effects on modularity compared with when not including this optional step.

The majority of generated candidates achieve a modularity score higher than the baseline threshold. However, the higher the candidate size, the lower the modularity is. Table 1 lists the protein group candidates for sizes from 2 to 10 achieving the highest modularity. Please note that the names of the proteins are the names of the corresponding genes approved by the HUGO Gene Nomenclature Committee (HGNC).

Large numbers of candidates are isolates (*i.e.*, consist of a single feature, or protein). From a purely biological point of view, one protein is not likely to be as relevant as a group of proteins because proteins tend to act in concerted fashion. Thus, it is rather difficult to make biological sense from information derived from isolates. It generally makes more sense to interpret the data of groups of proteins acting within a certain biological pathway.

Candidates are full of redundancy, especially the stronger ones. There are subsets of features that occur in more than one candidate. For example, `{GINS3, GINS4}`, `{GINS1, GINS2, GINS3}`, and `{GINS1, GINS2, GINS3, GINS4}` are on the list of candidates. In spite of the fact that modularity is inversely proportional to size, it might be desirable to give priority to large candidates when deciding on combinations of two candidates or more. It is worth remembering that a small candidate will not hold as much information as a large candidate.

We made a performance evaluation of the candidates' ability to cluster the samples, one candidate at a time. We applied Spectral Clustering on a sample graph that is generated using the samples as nodes and weight of edges is computing following Eq. (4) using only the proteins mentioned in the candidate features. Afterwards, we cluster the graph into two groups. Figure 5 shows the frequency of the size of the smallest cluster that we obtain when clustering the samples using the generated candidates. Clearly, the majority of candidates tend to split the samples in half. One would expect to see a small cluster with much lower or higher protein content, which suggests there exists a breast cancer subtype. However, we suggest that the combination of two candidates or more creates a fine-grain classification of breast cancer tumors.

There is a strong correlation between features and thus between candidates. While Figure 6 shows the correlation coefficient matrix of the 9,995 features, Figure 7 shows that there is a correlation between the candidates that are shown in Table 1. This is why, it is important to balance modularity and correlation, or candidates will overlap, and as a result it will not be possible to recognize fine-grain breast cancer subtypes.

It is possible to create a *co-occurrence graph*, i.e., H, which is used to encode the co-occurrence of proteins in candidates. After we select the top ten candidates for size from

Table 1
Top Candidate for Size from 2 to 10

Candidate	Modularity
SCGB1D2, SCGB2A2	0.22747956
GINS1, GINS2, GINS3	0.1854362
GINS1, GINS2, GINS3, CENPI	0.16841497
GINS1, GINS2, GINS3, STRA13, CENPI	0.15569982
CENPU, KIAA0101, NUSAP1, PBK, RRM2, TOP2A	0.15063133
FANCI, GINS1, GINS2, GINS3, STRA13, DTL, CENPI	0.13949169
FANCI, GINS1, GINS2, GINS3, KNSTRN, STRA13, DTL, CENPI	0.13492465
FANCI, GINS1, KNSTRN, KPNA2, SHCBP1, SKA1, UBE2T, DTL, SKA3	0.13174092
CENPU, FANCI, GINS1, GINS2, GINS3, KNSTRN, KPNA2, UBE2T, STRA13, DTL	0.13741883

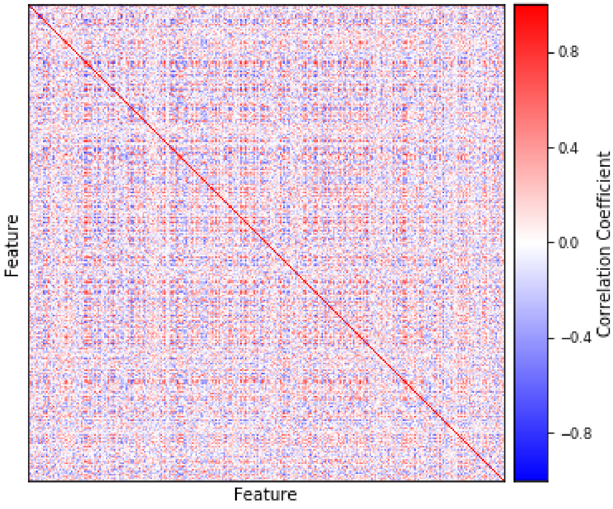

Figure 6: Correlation Coefficient Matrix

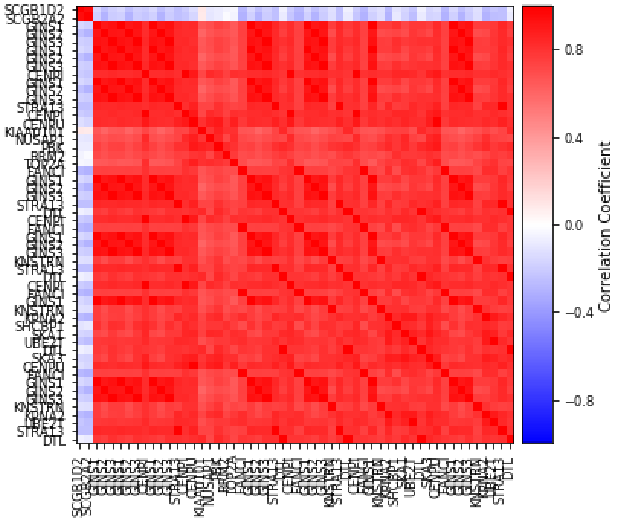

Figure 7: There is a correlation between the top candidates.

2 to 10, we add a node per protein present. There is an edge between u and v if these proteins co-occur at least once – in other words, if they are in the same candidate once or more. Also, the weight of the edge between protein u and protein v is the modularity score of the best candidate in which both u and v are present. H is not guaranteed to be connected. Indeed, there are 11 connected components in this graph, which are interpreted in Section 6. Figure 8 and Figure 9 show the largest and the second-largest connected component of the co-occurrence graph, respectively. Please note that node size is proportional to node frequency.

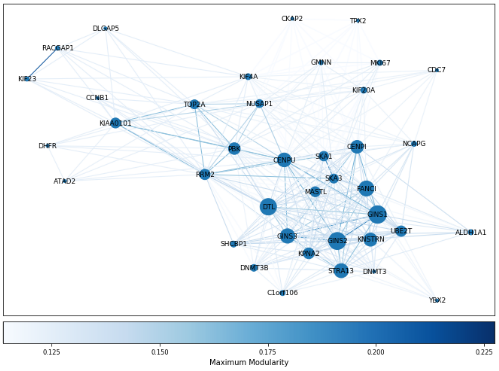

Figure 8: Largest Connected Component of H.

6. Biological Interpretation and Discussion

The candidates/protein clusters that showed top scoring modularities recapitulated many of the cellular phenotypes characteristic of cancer cells.

In the co-occurrence graph, there are 4 proteins out of 37 that find correspondence in PAM50.

In H, the largest connected component, which is composed of approximately 30 proteins, showed great internal consistency in its proteins with respect to their quantitative pattern across the tumor cohort, and reflected one of the most well known hallmarks of cancer, proliferation. Tumors can generally be divided into fast-growing, *i.e.*, proliferative, or slow-growing. This subdivision somewhat relates to the PAM50-based classification since most basal-like tumors tend to be highly proliferative, whereas most Luminal A and normal-like tumors tend to be slow-growing, but there is a mixed picture among the HER2 and Luminal B

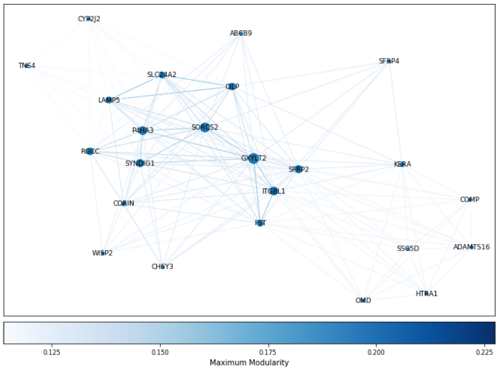

Figure 9: Second-Largest Connected Component of H

groups. At cellular level, several processes are conducive of this hallmark, and our current analysis highlighted local protein groups within the largest connected component (Figure 8). Thus, RRM2 and DHFR are enzymes carrying out nucleotide synthesis, required for DNA replication, which is a process initiated by the GINS complex (GINS1, GINS2, GINS3). And since DNA replication is a sensitive step, in which DNA is exposed to damage, one also observes a number of DNA repair proteins elevated in highly proliferative samples, such as KIAA0101 (a.k.a. PCLAF), UBE2T, DTL, and FANCI. The latter protein also assists in maintaining chromosomal stability, since the process of duplicating chromosomes is dangerous for their integrity. Also important for DNA and chromosomal integrity during this mitotic process are TOP2A, NCAPG and MKI67. Mitosis is the process of cell division/duplication. Many proteins that promote cell cycle progression into mitotic state are also elevated (in highly proliferative samples), such as CCNB1, DLGAP5, PBK and ATAD2. This latter one is particularly interesting since it mediates estrogen-induced proliferation, estrogen being a fundamental hormone for breast tissue. Many of the proteins involved in the actual structure of the mitotic spindle (namely NUSAP1, SHCBP1, KNSTRN, RACGAP1, KIF4A, and KIF23), which supports and aligns the duplicated chromosomes during mitosis, are also found in the largest connected component. And finally, the actual driver proteins of cytokinesis, which motor the chromosomes to the opposing cell poles during cell division are also found here, SKA1, SKA3, KIF20A, CENPU, and CENPI (Figure 8).

The second-largest connected component of H, which is made of 20+ proteins as shown in Figure 9, pointed to a group of proteins involved in extracellular matrix (ECM) homeostasis, including serine proteases and a metalloproteinase (CORIN, HTRA1, ADAMTS16) which cleave and trim substrates at the cellular surface, as well as proteins involved in glycosylation (GXYLT2, CHSY3). Other ECM related proteins are also found (KERA, ITGBL1, SSC5D, COMP). Intriguingly, several proteins with known retinal and corneal functions (HTRA1, KERA, ITGBL1, SLC24A2, and SFRP2) as well as proteins involved in synaptic plasticity in neurons (SORCs2, SYNDIG1, LAMP5) and proteins involved in collagen modulation (OMD, COMP, P4HA3) and cartilage development (CILP, TNS4) are found in this network. This suggests another familiar pattern in cancer cells taking place, that of expressing proteins with normal but rather specific functions in other tissues to the advantage of the tumor. It is notable that among these collagen and cartilage related proteins, several are involved in apoptosis (controlled cell death) regulation via caspase-3. Moreover, those proteins which suppress apoptosis (OMD, COMP) showed very similar quantitative patterns across the patient cohort, whereas the apoptosis promotor (TNS4) showed a rather opposing quantitative pattern. Finally, three proteins involved in Wnt signalling are also found here (SFRP2, SFRP4, and WISP2). Wnt signaling transduces signals of paracrine (i.e. from neighbouring cells) or autocrine origin into alteration of gene expression in the cell. Wnt signaling normally has functions in embryonic development and adult tissue regeneration, but it is well known be appropriated by cancer cells to their advantage. This ECM component had a pattern of expression across the tumor cohort almost opposite to the proliferation connected component, with predominantly high levels in Luminal A patients and low levels in Basal-like patients, and a mixed picture for the remaining subtypes.

Finally, a note on the enzymes chymase (CMA1) and carboxypeptidase CPA3. These are normally only expressed by mast cells (a type of immune system cell), the pair here shows an almost identical cross tumor expression pattern. This suggests that possibly the tumors with high levels of CMA1/CPA3 have more infiltration of this type of immune system cells. This property does not relate to the classic PAM50 classification system with several patients from the different subtypes showing high CMA1/CPA3 levels. This information about the type of immune system infiltration can potentially have decisive impact on the tumor behavior, including how it would respond to treatment, particularly in regards to the latest wave of cancer therapy, immunotherapy.

7. Future Work

We would like to explore further possibilities to enhance the methods adopted in our proposed unsupervised dimensionality reduction pipeline. For example, branch-and-bound algorithm can by used as a systematic approach to enumerate over the possible candidate solutions. Particularly, this algorithm represents the the search space as a tree with possible solutions and explores branches of this tree that maximize the objective function. Furthermore, we would like to incorporate different scoring functions in addition to modularity for the candidate evaluation step. Lastly, the majority of candidates are small and the high-performance candidates are isolates. For this reason, it might be desirable to merge two smaller candidates into one larger candidate and check if this merge brings any gain in terms of modularity. Accordingly, applying forward stepwise selection approach by adding independent features one at a time might be promising.

Since it is not certain that we find anomalies that might actually indicate breast cancer, it was necessary to manually double-check the results of the experiments against what we know about the identified candidate proteins. Despite overlapping, the large number of candidates justifies offering an automatic interpreter that supplements the results of the experiments. We plan to augment them by including external data sets such as Gene Ontology (GO) [1, 5]. It might be helpful to add the information about the roles of the proteins inside the cells, so that we can provide a brief overview of which biological processes are affected by the extracted anomalies. It might be helpful on a theoretical level because it could help indicate proteins worth examining in ad-hoc laboratory experiments.

8. Conclusion

We hereby propose a three-step wrapper method for the discovery of connected protein networks underlying particular molecular and cellular processes which characterize distinct behaviors in tumors in a manner agnostic/independent of the current PAM50-based breast cancer classification. Current patient stratification procedures for treatment assignment are often inadequate with many heterogeneous responses to a treatment within the same subtype. This is because the current classification is still too coarse, with many biological processes important for response outcomes not yet pinpointed.

By reducing the scope of investigation to a few sets of proteins at a time, it is possible to highlight the differences between the samples. For this reason, we generate candidates by splitting the feature set. In order to take advantage of Spectral Clustering, it was necessary to convert the data to a feature graph where the affinity between the proteins drives the splitting process. Large numbers of candidates were generated and it was necessary to evaluate them according to how good they performed on the task of clustering the samples. Candidates that are better than the state-of-the-art classification of breast cancer cells in terms of modularity find a correspondence in what we know about the functions of the these proteins.

Work has been done to extend the classification of breast cancer cells based on PAM50. Our new wrapper method is novel because it adopts a different strategy that is not biased because it is led exclusively by the cancer data. While some highlighted proteins are already known to relate to breast cancer, some are new and we believe it is worth examining them in depth. It is necessary to double-check that the proteins in the strong candidates are actually playing a role in the onset of the disease, and we performed an analysis at the theoretical level, manually. This is not scalable, but it serves the purpose of demonstrating the potential of the application of feature selection to breast cancer classification.

Relative to domain expertise-agnosticism, our method does not depend on a body of specialist knowledge, and consequently it is independent from our current comprehension of cancer biology. The first option of *candidate generation* analyzes the breast cancer data only, whereas the second option expands the search space by importing data from CORUM. We focused on CORUM, but it is easy to integrate other external data sets. Even if addition of external data might seem contradictory with the domain expertise–agnosticism requirement, we argue that no knowledge of the data, apart from its structure, is needed to take advantage of it and furthermore the external data is specific to proteins but not to cancer. However, it is worth remembering that the quality of the candidates of the second option depends on the quality of the external data, since these data sets are likely to be incomplete or biased.

As we have already mentioned, *feature filtering* is optional. It was added at a later time to deal with the space and time complexity of the algorithm caused by the large number of candidates generated. While it is optional, we believe it is worth losing some candidates in return for a speeding up of the pipeline. Indeed, we have seen a marked decrease in the number of candidates, but at the same time, the range of modularity has not been affected.

The main goal of this work is to help biologists with the identification of the markers of cancer types and subtypes. There are various treatments for cancer, but the fundamental idea is to treat each patient appropriately by adapting the treatment to the particular tumor phenotypes, which are essentially defined by the amount of individual proteins in the cancer cells. Course of treatment selection is based on the state-of-the-art classification of breast cancer tumors, but this is currently based on data obtained at the transcript (mRNA) level. Drugs are designed to target particular proteins, not their mRNA precursors. Thus, there is still much room for improvement. The study and treatment of cancer will undoubtedly benefit from the study and analysis of complex and comprehensive cancer proteomics data sets.

References

[1] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, Joel E.and Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25–29.

[2] Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is "nearest neighbor" meaningful?, in: International conference on database theory, Springer. pp. 217–235.

[3] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008. URL: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008, doi:10.1088/1742-5468/2008/10/p10008.

[4] Chao, G., Luo, Y., Ding, W., 2019. Recent Advances in Supervised Dimension Reduction: A Survey. Machine Learning and Knowledge Extraction 1, 341–358.

[5] Consortium, T.G.O., 2018. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research 47, D330–D338. URL: https://doi.org/10.1093/nar/gky1055, doi:10.1093/nar/gky1055.

[6] Cunningham, P., 2008. Dimension Reduction, in: Cord, M., Cun-

ningham, P. (Eds.), Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval. Springer, Berlin, Heidelberg. chapter 4, pp. 91–112. URL: https://doi.org/10.1007/978-3-540-75171-7_4, doi:10.1007/978-3-540-75171-7_4.

[7] Dias, K., Dvorkin-Gheva, A., Hallett, R.M., Wu, Y., Hassell, J., Pond, G.R., Levine, M., Whelan, T., Bane, A.L., 2017. Claudin-Low Breast Cancer; Clinical & Pathological Characteristics. PLoS ONE 12.

[8] Duval, B., Hao, J.K., Hernandez Hernandez, J.C., 2009. A Memetic Algorithm for Gene Selection and Molecular Classification of Cancer, in: Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, ACM, New York, NY, USA. pp. 201–208. URL: http://doi.acm.org/10.1145/1569901.1569930, doi:10.1145/1569901.1569930.

[9] Ellis, M.J., Gillette, M., Carr, S.A., Paulovich, A.G., Smith, R.D., Rodland, K.K., Townsend, R.R., Kinsinger, C., Mesri, M., Rodriguez, H., et al., 2013. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. Cancer Discovery 3, 1108–1112.

[10] Fortunato, S., 2010. Community detection in graphs. Physics Reports 486, 75–174. URL: http://www.sciencedirect.com/science/article/pii/S0370157309002841, doi:https://doi.org/10.1016/j.physrep.2009.11.002.

[11] Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., Ruepp, A., 2018. CORUM: the comprehensive resource of mammalian protein complexes—2019. Nucleic Acids Research 47, D559–D563. URL: https://doi.org/10.1093/nar/gky973, doi:10.1093/nar/gky973.

[12] Ha, M.J., Banerjee, S., Akbani, R., Liang, H., Mills, G.B., Do, K.A., Baladandayuthapani, V., 2018. Personalized Integrated Network Modeling of the Cancer Proteome Atlas. Scientific Reports 8.

[13] Huang, S., Cai, N., Pacheco, P.P., Narrandes, S., Wang, Y., Xu, W., 2018. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics & Proteomics 15, 41–51.

[14] Hutson, M., 2018. Has artificial intelligence become alchemy? Science 360, 478–478. URL: https://science.sciencemag.org/content/360/6388/478, doi:10.1126/science.360.6388.478, arXiv:https://science.sciencemag.org/content/360/6388/478.full.pdf.

[15] Johansson, H.J., Socciarelli, F., Vacanti, N.M., Haugen, M.H., Zhu, Y., Siavelis, I., Fernandez-Woodbridge, A., Aure, M.R., Sennblad, B., Vesterlund, M., Branca, R.M., Orre, L.M., Huss, M., Fredlund, E., Beraki, E., Garred, Ø., Boekel, J., Sauer, T., Zhao, W., Nord, S., Höglander, E.K., Jans, D.C., Brismar, H., Haukaas, T.H., Bathen, T.F., Schlichting, E., Naume, B., Geisler, J., Hofvind, S., Engebråten, O., Geitvik, G.A., Langerød, A., Kåresen, R., Mælandsmo, G.M., Sørlie, T., Skjerven, H.K., Park, D., Hartman-Johnsen, O.J., Luders, T., Borgen, E., Kristensen, V.N., Russnes, H.G., Lingjærde, O.C., Mills, G.B., Sahlberg, K.K., Børresen-Dale, A.L., Lehtiö, J., (OSBREAC), C.O.B.C.R.C., 2019. Breast cancer quantitative proteome and proteogenomic landscape. Nature Communications 10, 1600. URL: https://doi.org/10.1038/s41467-019-09018-y, doi:10.1038/s41467-019-09018-y.

[16] Liu, H., Yu, L., 2005. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering 17, 491–502. doi:10.1109/TKDE.2005.66.

[17] Mertins, P., Mani, D.R., et al., 2016. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 55–62.

[18] Newman, M.E.J., 2001. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E 64, 16132. URL: https://link.aps.org/doi/10.1103/PhysRevE.64.016132, doi:10.1103/PhysRevE.64.016132.

[19] Newman, M.E.J., 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103, 8577. URL: http://www.pnas.org/content/103/23/8577.abstract, doi:10.1073/pnas.0601602103.

[20] Ng, A.Y., Jordan, M.I., 2002. On Spectral Clustering: Analysis and an algorithm, in: Advances in Neural Information Processing Systems, pp. 849–856.

[21] Peng, S., Xu, Q., Ling, X.B., Peng, X., Du, W., Chen, L., 2003. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Letters 555, 358–362. URL: https://doi.org/10.1016/S0014-5793(03)01275-4, doi:10.1016/S0014-5793(03)01275-4.

[22] Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S.X., Lønning, P.E., Børresen-Dale, A.L., Brown, P.O., Botstein, D., 2000. Molecular portraits of human breast tumours. Nature 406, 747–752. URL: https://doi.org/10.1038/35021093, doi:10.1038/35021093.

[23] Şenbabaoğlu, Y., Sümer, S.O., Sanchez-Vega, F., Bemis, D., Ciriello, G., Schultz, N., Sander, C., 2016. A multi-method approach for proteomic network inference in 11 human cancers. PLoS computational biology 12, e1004765.

[24] Senkus, E., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rutgers, E., Zackrisson, S., Cardoso, F., 2015. Primary breast cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. Annals of oncology 26, v8–v30.

[25] Tukey, J.W., 1977. Exploratory Data Analysis.

[26] Tyanova, S., Albrechtsen, R., Kronqvist, P., Cox, J., Mann, M., Geiger, T., 2016. Proteomic maps of breast cancer subtypes. Nature Communications 7. URL: https://doi.org/10.1038/ncomms10259.

[27] Yanovich, G., Agmon, H., Harel, M., Sonnenblick, A., Peretz, T., Geiger, T., 2018. Clinical Proteomics of Breast Cancer Reveals a Novel Layer of Breast Cancer Classification. Cancer Research 78, 6001–6010.