Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

NAOMI: a network-based web application for matching cancer patients by integrating multiple data types

Simon Plovyt*, Milan Bracke*, Martin Misonne*, Mathijs Deprez*, Lieven P.C. Verbeke[1]

[1]Department of Information Technology, Ghent University—iMinds

*Contributed equally to this work.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genome diseases, such as cancer, are often very variable on the molecular level. Although genomic data of patients are becoming more and more available, the assignment of an effective treatment remains very challenging because of the abundance of unknown features. Matching patients on the molecular level in order to apply experience from previous cases on new cases has the potential to take medicine a big step forward towards the ultimate goal of treating cancer more effectively.
Results: NAOMI, a network-based algorithm was developed for specialists to access and upload the genomic information of new patients through a web application. NAOMI accurately returns the most similar database patients to allow for inference to the new patient based on past experiences, by integrating different data types. Validation tests to predict cancer subtypes return an accuracy of 78%.
Availability: The web application is available on https://github.ugent.be/mfbracke/design-project3.
Contact: lieven.verbeke@intec.ugent.be
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Each year, 14 million new cases of cancer are diagnosed and 8.2 million deaths are recorded worldwide (Ferlay *et al.*, 2013). We believe that these numbers should not be the standard. With the recent surge in personalised treatment, survival rates have increased and can be increased even further. To set up this treatment, knowledge of tumour types and subtypes is required, but this is obviously very challenging as cancer is a heterogeneous disease (Polyak, 2011). Indeed, many tumours behave in a similar way so patients might receive identical treatment. Although they may appear the same clinically, their alterations on the molecular level may be totally different. Therefore, the hardest and most critical part for doctors remains assigning the right treatment to his or her patients. Nowadays, therapies are usually based on the matching of certain biomarkers (Coppedè *et al.*, 2014) They consist of a small set of genes that have a high chance of being mutated in tumours. The downside of this approach is that tumours that should be treated the same way, are sometimes treated totally differently based on a single biomarker mismatch. Similarly, cancerous tissues that coincidentally share the same biomarkers may be treated identically although they are different on the molecular level. In this paper, we present our web application, named NAOMI (Network-based Application using Oncological data for Matching and Inferring). The method has the objective of reducing these false negatives and false positives using a network as a base, which represents the unique molecular details of each patient. Although the cost of such information (e.g., derived from whole tumour genomes) currently is prohibitively large, it is getting continuously lower and the data necessary to accurately predict molecular subtypes are progressively becoming available (NIH statistics, 2017). Many algorithms exist to stratify patients into meaningful classes, but only a few integrate all the different genomic data types that are available for each tumour (Hofree *et al.*, 2013, Ronglai et al., 2009, Shen *et al.*, 2012). Algorithms that do use networks don't stratify patients, but use them in a different way e.g. finding driver pathways for subtypes. (Le Van *et al.*, 2016, Verbeke *et al.*, 2015). From the few integrative algorithms that exist, none use networks in their approach (Wang et al., 2014). Also, the renowned IBM Watson for clinical trial matching does not focus on networks (ibm.com). Networks are highly advantageous in this approach, as they allow to integrate easily different data types. Secondly, they make the connection of upstream disturbances with downstream effects possible and thirdly, networks allow to match aberrant events that are scattered in a pathway or a local neighbourhood. This is important, because once a pathway is altered, it is less likely that another alteration will occur due to mutual exclusivity (Ciriello *et al*., 2011). Consequently, tumours of two patients with mutations in different genes, but in the same pathway, may be closely related. Also, mutations that occur in only few tumours, but again in the same pathway, are still strongly connected, in contrast to e.g. p53 which is mutated in many tumours.

To aid specialists in assigning a correct treatment to (cancer) patients, we have developed NAOMI to use a network that integrates different data types in order to find similar patients by measuring connectivity. Because we imagine our target users (medical practitioners) to require timely answers, the design was chosen to be as user-friendly as possible without requiring any technical knowledge. Users can simply upload the individual patient data (for example, but not limited to, mutational variants or differentially expressed genes) and are returned a list of similar patients. By means of guilt-by-association, properties of the patient under investigation and even suggestions towards the optimal therapy can be inferred from the best matching patients. Our motivation behind this project is the strong belief that assigning the correct treatment can be done based on past experiences, which are often disregarded on a large scale in a clinical setting.

2 NAOMI

At the backend of our web application, NAOMI (Network-based Application using Oncological data for Matching and Inferring) selects and returns the most similar patients. This algorithm is centered around the use of a network as a representation of the information held by database patients, in combination with prior knowledge. Such a network representation allows for the calculation of distance measures in a straightforward way. Several of these distance measures are available (Fouss *et al.*, 2012), such as the Laplacian exponential diffusion kernel or more commonly, a random walker, and can be applied to determine similar patients with considerable accuracy. We will elaborate on the proposed method using examples and concepts drawn from the data on which the method will eventually be evaluated. These data are described in detail in the Data section.

2.1 The network

The backbone of the algorithm is the network. Networks allow to integrate different data types, they make the connection of upstream disturbances with downstream effects possible and allow to match aberrant events that are scattered in a pathway or a local neighbourhood. Genomic data of the database patients are integrated together around the patient (Figure 1), by connecting the patient node to the nodes of mutated, differentially expressed, hypo- or hyper-methylated genes and genes with substantially different copy numbers. Note that data types can be omitted, or can be added without altering the main ideas presented here. The only limitation is that all data should be binary, where a 0 represents a normal state of a gene for a particular data type, and a 1 represents an aberrant state, e.g. a mutation, differential expression, … Thus, close patients will be indirectly connected by connecting to the same aberrant genes or to closely related aberrant genes. The big advantage of this network-based approach is that it allows to also interconnect genes that are known to be in the same pathway. In a biological setting, e.g. a mutation can affect related genes or even entire pathways and this we attempt to simulate. Therefore, such connections were derived from databases such as REACTOME (Croft *et al.*, 2011), KEGG (Kanehisa *et al.*, 2000), STRING (Szklarczyk *et al.*, 2011), BioGRID (Stark *et al., 2006)* , … to provide a supportive net under the database patients. This implementation adds useful prior knowledge to perform more accurate patient matching. Another key property of this network is that it is undirected. This is based on the intuitive reasoning that we only want to measure connectivity and thus directions are unnecessary. Furthermore, the network is unweighted, because there is no clear way to weigh the edges. However, this implies that continuous data such as expression values or methylation values must be made binary. Subsequently, the input patient will be added in a similar manner. Finally, the corresponding adjacency matrix is computed to efficiently calculate distance.

2.2 Scoring connectivity

Three similarity metrics were evaluated here: a mathematical implementation of the random walker with restart, the Laplacian exponential diffusion kernel and an actual simulated random walker (ref. Materials and methods). As a result, the connectivity score can be obtained between any two nodes and thus also between patients. The highest scoring patients are of interest and will be selected.

2.3 Use of the web application

The web application was designed to be easy to use and to not require any technical knowledge. Besides uploading the data, most parameters are chosen automatically or set to the default value and results are merely one click away. Next to the most similar patients, graphs and charts are returned to present the data in a pleasant manner and to allow for inferring characteristics in minimal time for maximal efficiency.

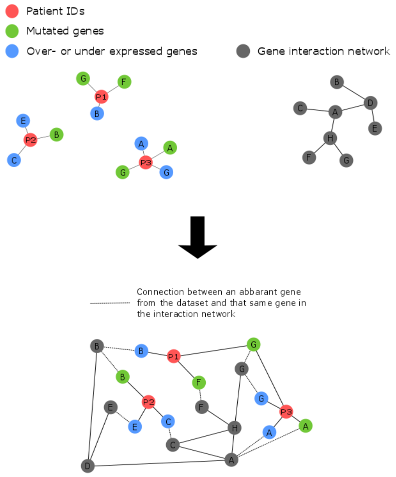

Patient IDs
Mutated genes
Over- or under expressed genes
Gene interaction network
P1
P2
P3
Connection between an abbarant gene from the dataset and that same gene in the interaction network

Figure 1: Diagram of the construction of the global network. This simplified representation only shows the combination of three data types; mutation data and expression data, in combination with the REACTOME interaction network. The red nodes represent the patients, whereas green and blue nodes are mutated versions and differentially expressed versions of genes, respectively. The genes found in the interaction network of choice are coloured grey. These patient nodes are first linked to their corresponding aberrations. Then, the aberrations are linked with their corresponding gene in the interaction network (prior knowledge).

3 Materials and methods

3.1 The data

The used dataset was a small, but complete set of breast cancer patients (BRCA set, obtained from TCGA and described in Cancer Genome Atlas Network, 2012) that consists of both expression values and simple mutation data. Because this dataset is relatively small (463 patients), this set is used for testing and validation purposes.

Mutations in the small BRCA dataset were processed with the MUTSIG method (Lawrence *et al.*, 2012) and come in binary format without variant allele frequency (VAF) reliability score. When a VAF score is not given, we assign a value of 1 to make sure the genes are not removed. Expression data are continuous and thus has to be made binary (see further). In regards to clinical data, for the small BRCA dataset only the tumour subtypes (PAM50) were available and can be used as a metric for validation.

3.2 Binarization

In order to represent aberrant genes in an unweighted network, the problem arose to determine whether a gene is considered abnormal (1) or not (0). For mutational data, which is binary by nature, no processing is required besides filtering unreliable calls based on VAF score (see further). However, for continuous data such as expression values, copy number and methylation scores, a different approach is needed. To describe the binarization process, the example of expression data is used.

To determine if a gene is behaving abnormally, we need to define what is considered normal. For every gene, all expression values are collected to build the distribution. Based on this distribution, the median is computed. Now, symmetrical top and bottom percentiles (defined by a user chosen parameter, e.g. 5^{th} and 95^{th} percentiles (5% and 95%)) are evaluated. The difference in value between the bottom percentile and the median is compared to the difference between the top percentile and the median. The smallest difference is mirrored around the median and these thresholds are used to determine differential expression (Figure 2).

Finally, when constructing the network, the patients will be connected to nodes of their aberrant genes.

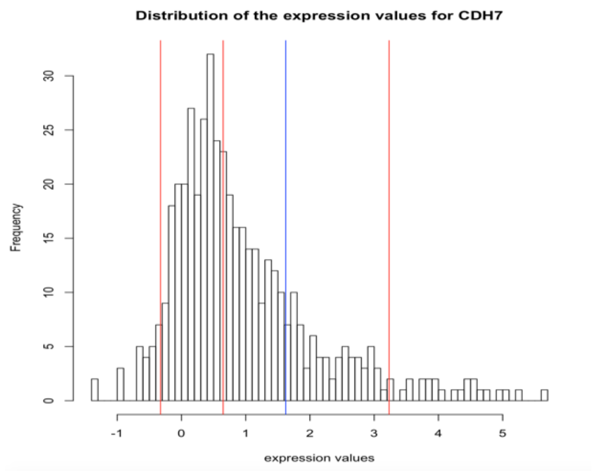


Figure 2: Distribution of the gene expression values for a randomly chosen gene (CDH7 in this case). The red lines represent the 5th percentile (5%, left), the 50th percentile (median, center) and the 95th percentile (95%, right). Because the 5th percentile is closer to the median, the distance from this percentile to the median is mirrored and represented by the blue line. Everything to the right of the blue line will therefore be considered overexpressed (in this case) and assigned the value of 1 in the binarization process.

3.3 Basic algorithm

Before developing a more complex algorithm, we decided to implement a less sophisticated version that does not use networks as a baseline. This allowed us to quickly understand the challenge of cancer subtype identification. As expected, this algorithm was built fast and made testing the web application more convenient by simplifying the debugging process. Moreover, it only uses one type of data, which allowed working on a standard machine without encountering memory issues and tremendous computational time. The accuracy of this algorithm will function as a benchmark for the performance of the more complex approach.

The first abnormality that comes to mind when thinking of the reasons for cancer are mutations. Therefore, and for the sake of simplicity (binary by nature), we have decided to use this kind of data as a base line, while we could also use copy number variation, expression or region specific mutation data. However, some patients can share the same subtype of cancer even if they don't have any single mutated gene in common. Indeed, if the same pathway is mutated for both patients, equal phenotypes may be observed. For example, a perturbation in gene A may impact genes B and C downstream the pathway and result in the same tumour phenotype as if only B or C were altered. This is why we decided to represent the mutations of the patients on the pathway level and not on the gene level.

To represent the data, we built a vector for each patient consisting of one score for each pathway. The score represents the number of mutations present. Next, the basic algorithm calculates the distance (dissimilarity) between a new patient and each other database patient based on one of the following metrics: Euclidian distance, Manhattan distance or Jaccard distance. The smaller the distance between two patients, the more similar they are.

The difference in size of the pathways could make our results biased towards larger pathways. Assuming that the average length per gene is constant across the pathways, the larger a pathway is, the more likely it is that a mutation occurs. Thus, two patients can be considered more similar just by random chance. For this reason, we divided each score by the length of its pathway as a way of normalizing the scores.

The subtype of the new patient is decided by considering the most represented amongst the similar patients. This will be discussed more extensively later.

3.4 Network construction and filtering

Following some basic data cleaning to remove incomplete entries, the adjacency matrix to represent the network can be created. The elements of this adjacency matrix indicate whether two nodes are direct neighbours (1) or not (0). This requires a collection of all nodes. First, the REACTOME interaction network was imported. Secondly, all database patients were imported and a node for every patient was added to the set. Moreover, every data type was checked and the underlying genes were added to the collection set after adding a prefix to indicate the source of the gene (mutated, differentially expressed, interaction network, …). At this step, the user can decide whether all genes are collected or only the genes that are also present in the interaction network and thus supported by prior knowledge (default).

Now, based on the number of all nodes, the empty adjacency matrix was created. The first step in filling this matrix is adding the interactions present in the REACTOME network. In practice, this means connecting the interaction network nodes to their interaction partners, by adding the value of 1 in the corresponding column and row of the matrix. The network is undirected and thus the connections were made symmetrically. To create a supportive net for the database patients, every interaction network gene needed to be connected to other variations of the same gene, such as the mutated version, differentially expressed version, abnormal copy number version and also the version with deviating methylation patterns. Next, the database patients were connected to their deviating genes if this gene meets the criteria; a large enough variant allele frequency (VAF) score for mutated genes or the confirmation of being substantially different (1 after the binarization process) for continuous data sources.

3.5 Calculating scores

To derive similarity from the adjacency matrix representation of the global network, a suited similarity measure is required. Intuitively, shortest paths calculations are a possibility. However, this has been shown to underperform, especially when data is incomplete or qualitative (Suthram *et al*., 2008; Verbeke *et al*., 2013).

Kernels based on graph nodes generally tend to perform better (Verbeke *et al*., 2013; Qi *et al*., 2008; Nitsch *et al*., 2010; Lavi *et al*., 2012) and previous preliminary research (Verbeke *et al*., 2015) suggested that the Laplacian Exponential Diffusion kernel yields stable results.

Laplacian Exponential Diffusion (LED) kernel

The available implementation of the LED kernel computes the Laplacian matrix from the adjacency matrix A.

$$\mathbf{L = laplacian(A)} \tag{1}$$

Next, the global similarity matrix K is found by the following:

$$\mathbf{K_{LED} = exp(-\alpha L)} \tag{2}$$

Here, α is the parameter that quantifies how much one abnormal gene influences its surrounding genes, whereas exp is the matrix exponential. α must be carefully chosen as it controls the locality of the gene effect. If α is too low, the nodes do not substantially influence neighbouring nodes, which nullifies the advantage of using a network. However, when α is chosen too high, one node influences too many neighbours, which is biologically speaking not plausible. This can be interpreted as the amount of time one allows ink to diffuse when placing a drop on top of a node.

Random Walker

Both a mathematical matrix implementation as a simulated model of the random walker concept are made available. For the matrix implementation, the entire global similarity matrix is computed as follows:

$$\mathbf{K_{RW} = (D - \beta A)^{-1} * D} \tag{3}$$

Here, D is the degree matrix and β is the reset chance for the random walker to return to the node of origin.

As both the LED kernel and the full matrix implementation of the random walker are performed on the entire adjacency matrix, the required computational time is no longer negligible. To facilitate testing, we therefore implemented a simulated version of the random walker with restart. This operation starts at the input patient node and simulates a traveller visiting neighbouring nodes. Every neighbour is randomly chosen, while taking into account a chance β to return back to the input patient node. This process is local around the input patient and is therefore not subject to the size of the adjacency matrix. A large amount of iterations is performed while counting the number each neighbouring patient node was visited.

3.6 Parameters

Mutation reliability threshold

This parameter determines at which reliability (variant allele frequency, VAF) score mutations are filtered. All mutations that score below the cutoff value are too unreliable to be transferred into the network. A good threshold is found to be at a value of about 0.4 after a series of optimization tests. This is also low enough to prevent dilution of the mutations by other data types in the network.

Percentile of aberrant genes in context of binarization

With a default value of 1, this parameter decides what percentiles are considered to be aberrant in a distribution of continuous values. In the context of gene expression, this value of 1 is equal to the genes in the <1st and >99th percentiles. The percentile closest to the value of the median (50%) is mirrored to find the actual threshold values used to decide whether a gene is differentially expressed (DE) (1) or not (0). If this parameter is too big, too many genes will be considered to be DE when they are not, with consequences such as low prediction accuracy. Choosing a value too low will reduce the overall effect and presence of expression data in the global network.

Filtering genes by interaction network

This binary parameter (true by default) determines whether only gene nodes that are present and supported by the interaction network should be included in the global network.

Similarity metric parameters

Laplacian exponential diffusion kernel parameter α:

α in the LED kernel translates into the parameter describing the effect of one node on other nodes. Metaphorically speaking, when ink is placed on one node of the network, α is the time the ink gets to diffuse and to increase the colouring radius around the starting node. Choosing α to low restricts the effect of nodes on nearby nodes too much, nullifying the advantage of using a network. An exaggerated value for α increases the effect too drastically. The latter would transfer the aberrant gene effect to genes that are not related, which is not plausible in a biological context.

Good and stable values for α were suggested (Verbeke *et al*., 2015) and confirmed to be between 0.0001 and 0.05.

Random walker parameter β:

The chance for the random walker to reset back to the original position is given by parameter β. Defining β too high restricts the walker from exploring the neighbourhood, whereas a very low value of β allows the walker to visit too many patient nodes to find the closest. Optimising using validation accuracy has lead us to assigning 0.2 to β.

Number of similar patients

The quantity of most similar patients returned. Naturally, this parameter has no effect on determining these similar patients. However, when inferring patient characteristics, the number of patients to infer from is critical. In the context of validating the algorithm, this parameter defaults to 7 when using the small BRCA dataset for validation.

One way to remove this parameter is the implementation of an automated way to extract the most important similar patients based on the score distributions (see Discussion).

3.7 Validation

To measure the performance of our algorithm and to estimate optimal values for the parameters, we resort to a cross validation approach. We have used the relatively small BRCA dataset in order to validate using the leave-one-out method. One database patient is left out, the adjacency matrix is created and the left out patient is added back as input patient. Next, similarity scores are computed and the best scoring patients are selected. The default setting for this small dataset is to extract the top 7 patients. The overrepresentation of subtypes in these top patients is assessed and the most represented subtype is taken as the predicted subtype and compared to the known subtype of the previously left out input patient to calculate accuracy.

3.8 Web application

Our application consists of a web frontend, using AngularJS (angularjs.org) and a backend written in Python 3.5. The backend has a REST API. The client application that consumes this web service provides the specialist with the ability to see a list of all the patients in the database, enter new patients, modify patient data and view similar patients using our algorithm. Every patient has a unique id, but to have an identifier the user can make use of the option to enter an alias for the patient. This alias can then be used to look up the patient in the database. To protect the privacy of the patients, their names are not used in the application. Not using the names is also applicable in real-life situations.

The backend was built using a layered architecture, as can be seen in the deployment diagram (Figure 3). For communication with the frontend we have an api layer. The api layer controls the application through the controllers' layer. Controllers keep track of the right domain layer objects and provide functions to safely change and request information about these objects. The domain layer contains the domain (oncology) logic. It's split into two modules. The state module contains models that represent the info we have, for example patient data. It also makes sure that these models are always in a consistent state, so we don't have impossible data. And the computation module contains the algorithm to compute similar patients.

The application is still a prototype, so modifications would have to be made before it can be used in real-life situations. The most important features that are missing are a database and a security layer. The database can be added by adding a persistence layer and connecting the domain.state module to this layer. Security can be implemented in different ways. An in-depth discussion would be beyond the scope of this paper. The application should also be tested more extensively before it can be used to treat patients.

3.9 Complexity

The time complexity of the algorithm is $O((g+p)^3)$ and the memory complexity is $O((g+p)^2)$, where g is the number of genes and p the number of patients in the matrix. This is because the inversion and the calculation of the matrix exponential of a large Laplacian matrix dominate the running time and memory consumption.

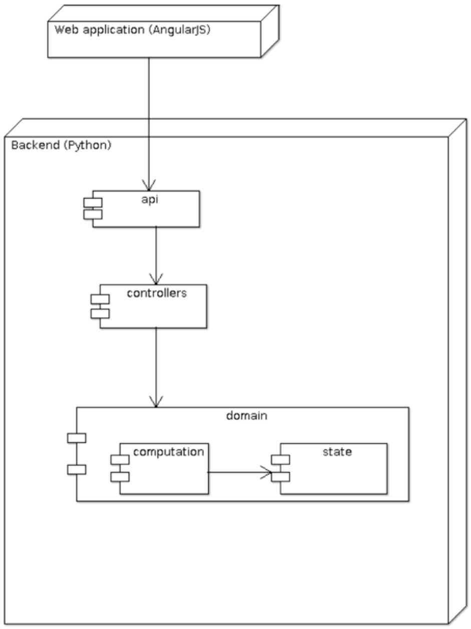


Figure 3: Deployment diagram of the application.

4 Results & Discussion

4.1 Basic algorithm

The accuracy obtained when assigning a subtype to a patient using the less sophisticated algorithm (using (normalized) pathway scores based on mutations only) on the BRCA dataset was 52.3% using Euclidian distance as metric. This is better than the accuracy obtained by assigning the subtype randomly, which is 30.6% (Figure 4).

4.2 NAOMI

NAOMI was put to the test with the breast cancer dataset of 463 patients by using the leave-one-out cross validation approach. The accuracy of the algorithm to predict the breast cancer subtype of the left-out patient, given the data of the 462 remaining patients, was evaluated over different settings (Figure 4).

Originally, the subtypes were clinically determined by a transcriptomics analysis. Therefore, feeding gene expression data to the algorithm is expected to drastically improve the results. Indeed, the accuracy when using only mutation data was 56% at best over the three different scoring metrics, which is only slightly better than the basic approach (about 52%). However, when integrating the mutation data together with the expression data, the accuracy jumped to 78% as the best result over the different scoring metrics, under the assumption that all clinically determined subtypes were assigned perfectly accurate.

When comparing the different scoring metrics, the random walker method (accuracy of 56%) slightly outperformed the Laplacian exponential diffusion kernel (accuracy of 52%). However, the simulated random walker quickly lost its leading position in the ranking when also integrating the expression data. The accuracy fell to about 30% and is therefore on the same level as random subtype assignment. The cause of this problem could be found in the fact that the expression data contains highly connected nodes. These nodes draw the simulated random walker towards some very 'expressive' patients and the walker tends to gets lost in the attractor cycles surrounding or containing some expressive patients. As a result, these patients were often considered 'similar', which incorrectly predicted the basal subtype in many cases.

The Laplacian exponential diffusion kernel method did not encounter the same problem. Instead, making use of this method increased the accuracy to a very decent 78% after some basic parameter optimisation. For information, an out-of-the-box machine learning model (implemented in python using scikit-learn and taking the best performing out-of-the-box classification model as standard, which was a random forest classifier) has an accuracy of up to 75% of prediction subtypes with the same binary data (Figure 4).

The remaining incorrect predictions are represented in a confusion matrix to draw conclusions for further research (Table 1). The dataset contains patients from 4 different subtypes of breast cancer: Basal-like, Luminal A, Luminal B and Her2. The accuracy for predicting the Basal-like and the Her2 subtype are near perfect, with 97% and 96% respectively, whereas the Luminal A and B subtypes are often confused with each other. In 11% of the cases were the subtype is Luminal A, it is wrongly predicted as Luminal B and in 34% of the cases were the actual subtype is Luminal B, it is wrongly predicted as Luminal A. These subtypes are of course quite similar and will require further optimisation of the algorithm to allow more reliable predictions.

Using NAOMI, the 10 most shared gene aberrations between correctly predicted patients of every subtype (Table 2) and the corresponding Venn-diagram representation (Figure 5) were extracted and identified. These genes are considered the most important to predict a subtype. It appears that some of those genes play a role in breast cancer (KCNJ3, PIK3CA, BMPR1B, ESR1, …) (Kammerer *et al.,* 2016; Young *et al.,* 2016; Sætrom *et al.*. 2009; Spoerke *et al.,* 2016). Also, the mutated gene TP53 tops the gene ranking of Her2 (Table 2) and is known to be characteristic of the Her2 subtype (Rath *et al.,* 2013). Therefore, this approach is also considered qualified to identify new biomarkers for less common types of cancer.

The predicted subtype is determined by the most represented subtype amongst the most similar patients. The default number of similar patients returned by the algorithm for the validation test was 7. As for the web-application itself, the user can select how many top patients to find returned. When using a dataset that contains more detailed properties (i.e. data from the ICGC database), such as tumour grade, patient survival rate or tissue type, inferring treatment or assigning other properties to the input patients is a lot more complicated than predicting the subtype. Therefore, our algorithm does not predict properties of the input patient, but simply returns the chosen amount of most similar patients and their corresponding properties. Inference can then be done by a medical specialist. This is an entirely different field of expertise and thus will not be discussed here. NAOMI solely aids the specialist by returning useful information represented in charts (Figure 6).

Table 1: Confusion matrix for the obtained prediction accuracy by NAOMI. This is the result of the leave-one-out cross validation approach on the BRCA dataset containing 463 breast cancer patients from TCGA.

	Confusion	Actual			
	Matrix	Basal	Lum A	Lum B	Her2
Predicted	Basal	97%	1%	2%	4%
	Lum A	1%	76%	34%	0%
	Lum B	0%	11%	56%	0%
	Her 2	2%	2%	8%	96%

Table 2: Top 10 shared genes between similar patients and input patients for correct predictions of every subtype. Mutated versions of genes are underlined, whereas the remaining genes are differentially expressed versions of the genes (not underlined). These results were obtained using NAOMI on the BRCA dataset containing 463 breast cancer patients.

LumA	LumB	Basal	Her2
BMPR1B	KCNJ3	B3GNT5	TP53
VIPR2	BMPR1B	TP53	PNMT
CA4	PPP1R1B	PSAT1	ESR1
GHRH	COL5A2	SCUBE2	ERBB2
GRIA2	COL11A1	GFRA1	FGFR4
ABCA9	INSM1	HIST1H1A	PSMD3
KCNJ3	DCN	KCNK5	CRYM
PIK3CA	MAT1A	MUC16	FABP6
LIN7A	LAMC2	IL12RB2	GJB1
GPD1	IL19	KCNG1	IYD

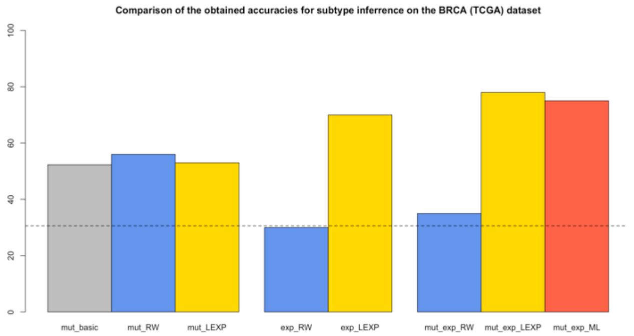


Figure 4: Comparison of the cross-validation accuracies (%) obtained by NAOMI for subtype inference on the BRCA dataset. Group 1 (left) uses only mutation data with the basic algorithm (52.3%), random walker (RW, 56%) (NAOMI) and Laplacian exponential diffusion (LEXP, 52%) kernel methods (NAOMI). The second group uses only the differentially expressed genes for the random walker method (30%) (NAOMI) and the LEXP method (75%) (NAOMI). The third group integrates both mutation and binary expression data for the random walker method (35%) (NAOMI), LEXP method (78%) (NAOMI) and an out-of-the-box random forest machine learning classifier (75%), respectively. The grey dashed line represents the accuracy of randomly assigning subtypes (30.6%).

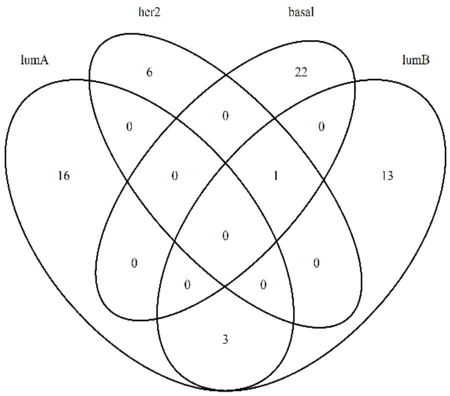


Figure 5: Venn diagram of the top genes shared between patients with a correctly classified subtype compared with other subtypes. Most top genes are specific for the subtype itself. However, correct Luminal A subtyped patients share some top genes with correct Luminal B subtyped patients. This is to be expected, as the Luminal A and Luminal B subtypes are quite similar, which also explains the high grade of confusion between both subtypes (Table 1).

Instead of a hard, user-defined cut-off for the top patients, we are currently investigating a more objective way to to determine the appropriate score threshold to use based on the score distributions of the collection of similar patients (see Figure 7 for sample score distributions of the result for three different input patients).

Other improvements can certainly be suggested on different levels and should be subject to further research to optimise NAOMI. For the network, we recognize it might be of interest to add weights representing the continuous data to the edges and develop an appropriate scoring function.

Secondly, improving the binarization function to more conveniently determine and capture the effect of aberrantly expressed genes, genes with different copy number, differently methylated genes or other continuous data in general. One way to improve this function might be to correct for false positives (false DE genes), depending on the distribution of the expression data points.

Furthermore, for real cases of treatment inference following patient matching, access to larger and more annotated datasets, such as data from the ICGC database, can be requested. For these genomic data, more elaborate information is often available, i.e. age, gender, treatment... even smoking history. This information is extremely useful when inferring characteristics for the input patient by extending from the found similar patients.

We intend to expand the algorithm to these more complex datasets, while also exploring the possibilities of the well performing machine learning approaches, with the ultimate goal of perfecting the algorithm and to more efficiently beat cancer.

5 Conclusion

We have developed a network-based web-application named NAOMI to reliably return similar patients from the TCGA database (BRCA breast cancer dataset) and their properties based on matching genomic data to an input patient. This algorithm has important applications in the medical field. It allows experts to use previous experiences concerning cancer therapy to infer suited treatments for new patients. Besides treatment purposes, other patient characteristics can also be predicted, such as tumour subtype, which was used as a way to validate our approach to an accuracy of 78%.

At the time of writing, this web application is intended for oncologic research. However, we will not limit NAOMI to this field in the future as we will attempt to generalize this approach to other genetic diseases, while expanding with different machine learning approaches. For example, adapting NAOMI to the case of autism which often comes with a spectrum of other disorders such as inflammatory bowel disease or muscular dystrophy (Nazeen *et al*., 2016). Using multiple genomic data sets may also help to better understand numerous illnesses such as diabetes (Bergholdt *et al*., 2007), Schizophrenia (Xiongjian *et al*., 2014) and other complex diseases (Sun and Hu, 2016). These examples emphasize the importance of further research on this topic.

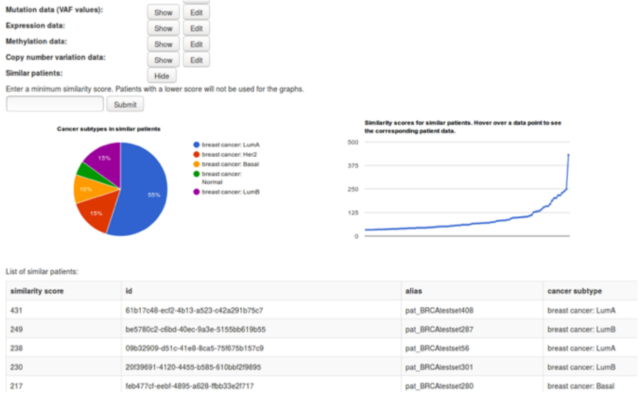
Acknowledgements

This work has been supported by Ghent University.

Funding

This work has been supported by Ghent University.

Conflict of Interest: none declared.



Enter a minimum similarity score. Patients with a lower score will not be used for the graphs.

List of similar patients:

similarity score	id	alias	cancer subtype
431	61b17c48-ecf2-4b13-a523-c42a291b75c7	pat_BRCAtestset408	breast cancer: LumA
249	be5780c2-c6bd-40ec-9a3e-5155bb619b55	pat_BRCAtestset287	breast cancer: LumB
238	09b32909-d51c-41e8-8ca5-75f675b157c9	pat_BRCAtestset56	breast cancer: LumA
230	20f39691-4120-4455-b585-610bbf2f9895	pat_BRCAtestset301	breast cancer: LumB
217	feb477cf-eebf-4895-a628-ffbb33e2f717	pat_BRCAtestset280	breast cancer: Basal

Figure 6: Screenshot of the first fully functional implementation of NAOMI.

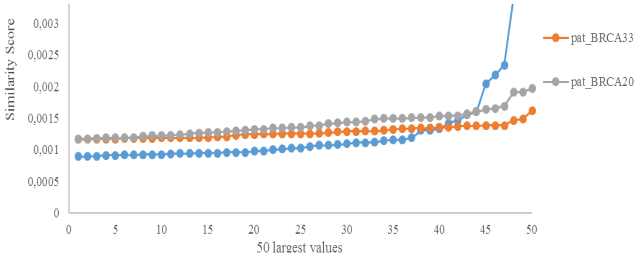


Figure 7: Distribution of the top 50 patient similarity scores for 3 random patients of the BRCA test set, using the Laplacian exponential diffusion kernel in NAOMI.

References

Bergholdt R., Størling Z. M., Lage K., Karlberg E. O., Olason P. I., Aalund M., Nerup J., Brunak S., Workman C. T. and Pociot F. (2007), Integrative analysis for finding genes and networks involved in diabetes and other complex diseases. *Genome Biol.*

Cancer Genome Atlas Network. (2012), Comprehensive molecular portraits of human breast tumours. *Nature*

Cho S. H., Jeon J. and Kim S. I. (2012), Personalized Medicine in Breast Cancer: A Systematic Review. *J Breast Cancer.*

Ciriello G., Cerami E., Sander C. and Schultz N. (2011), Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*.

Coppedè F, Lopomo A, Spisni R, Migliore L. (2014) Genetic and epigenetic biomarkers for diagnosis, prognosis and treatment of colorectal cancer. *World Journal of Gastroenterology*

Croft D, O'Kelly G, Wu G, et al. (2011), Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*

Ferlay J., Soerjomataram I.., Ervik M., Dikshit R., Eser S., Mathers C. et al (2012), GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11

Fouss F., , Francoisse K. , Yen L., Pirotte A. and Saerens M. (2012), An experimental investigation of kernels on graphs for collaborative recommendation and semi-supervised classification. *Neural Netw.*

Hofree M, Shen J. P., Carter H., Gross A. and Ideker T. (2013), Network-based stratification of tumor mutations. *Nature*.

Kammerer S, Jahn SW, Winter E, *et al.* (2016), Critical evaluation of KCNJ3 gene product detection in human breast cancer: mRNA in situ hybridisation is superior to immunohistochemistry. *Journal of Clinical Pathology.*

Kanehisa, M., & Goto, S. (2000), KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*

Lavi O., Dror G. and Shamir R. (2012), Network-induced classification kernels for gene expression profile analysis. *Comput Biol*.

Lawrence M. S. *et al.* (2013), Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.*

Le Van T., van Leeuwen M., Carolina Fierro A., De Maeyer D., Van den Eynden J., Verbeke L., De Raedt L., Marchal K. and Nijssen S. (2016), Simultaneous discovery of cancer subtypes and subtype features by molecular data integration. *Bioinformatics*.

Nitsch D., Gonçalves J. P., Ojeda F., de Moor B. and Moreau Y. (2010), Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*.

Qi Y., Suhail Y., Lin Y-y, Boeke J. D. and Bader J. S. (2008), Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res*.

Nazeen S., Palmer N. P., Berger B. and Kohane I. S. (2016), Integrative analysis of genetic data sets reveals a shared innate immune component in autism spectrum disorder and its co-morbidities. *Genome Biol.*

NIH statistics (2017) https://www.ncbi.nlm.nih.gov/genbank/statistics/

Polyak K. (2011), Heterogeneity in breast cancer. *The Journal of Clinical Investigation.*

Ronglai S., Adam B. O. and Marc L. (2009), Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Gen Biol.*

Sætrom, P. *et al.*, (2009), A risk variant in a miR-125b binding site in BMPR1B is associated with breast cancer pathogenesis. *Cancer Research.*

Shen R. *et al.* (2012), Integrative Subtype Discovery in Glioblastoma Using iCluster. *PLoS ONE*.

Spoerke, J. M. *et al.* (2016), Heterogeneity and clinical significance of ESR1 mutations in ER-positive metastatic breast cancer patients receiving fulvestrant. *Nature Communications*.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006), BioGRID: a general repository for interaction datasets. *Nucleic acids research*

Sun Y. V. and Hu Y. J. (2016), Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. *Adv Genet.*

Suthram S, Beyer A, Karp RM, Eldar Y and Ideker T. (2008), eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol Syst Biol*.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., … von Mering, C. (2011), The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*

Verbeke LPC, Cloots L, Demeester P, Fostier J and Marchal K. (2013), EPSILON: an eQTL prioritization framework using similarity measures derived from local networks. *Bioinformatics*.

Wang B., Mezlini A. M., Demir F., Fiume M., Tu Z., Brudno M., Haibe-Kains B. and Goldenberg A. (2014), Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods.*

Xiongjian L., Liang H., Leng H., Zhenwu L., Fang H., Roger T. and Lin G. (2014), Systematic Prioritization and Integrative Analysis of Copy Number Variations in Schizophrenia Reveal Key Schizophrenia Susceptibility Genes. *Schizophr Bull.*

Young C. D. *et al.* (2015). Activating PIK3CA Mutations Induce an Epidermal Growth Factor Receptor (EGFR)/Extracellular Signal-regulated Kinase (ERK) Paracrine Signaling Axis in Basal-like Breast Cancer. *Molecular & Cellular Proteomics.*