BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 4. 612–625
Advance Access published on 18 December 2012
doi:10.1093/bib/bbs083

Network information improves cancer outcome prediction

Janine Roy, Christof Winter, Zerrin Isik and Michael Schroeder

Submitted: 21st August 2012; Received (in revised form): 30th October 2012

Abstract

Disease progression in cancer can vary substantially between patients. Yet, patients often receive the same treatment. Recently, there has been much work on predicting disease progression and patient outcome variables from gene expression in order to personalize treatment options. Despite first diagnostic kits in the market, there are open problems such as the choice of random gene signatures or noisy expression data. One approach to deal with these two problems employs protein–protein interaction networks and ranks genes using the random surfer model of Google's PageRank algorithm. In this work, we created a benchmark dataset collection comprising 25 cancer outcome prediction datasets from literature and systematically evaluated the use of networks and a PageRank derivative, NetRank, for signature identification. We show that the NetRank performs significantly better than classical methods such as fold change or *t*-test. Despite an order of magnitude difference in network size, a regulatory and protein–protein interaction network perform equally well. Experimental evaluation on cancer outcome prediction in all of the 25 underlying datasets suggests that the network-based methodology identifies highly overlapping signatures over all cancer types, in contrast to classical methods that fail to identify highly common gene sets across the same cancer types. Integration of network information into gene expression analysis allows the identification of more reliable and accurate biomarkers and provides a deeper understanding of processes occurring in cancer development and progression.

Keywords: *network-based; outcome prediction; gene expression; PageRank; cancer biomarker*

INTRODUCTION

Cancer outcome prediction aims to address the problem of learning how to forecast disease progression from gene expression and thus allowing refined treatment options. A recent success in the area is the diagnostic kit MammaPrint®, which predicts metastasis formation in breast cancer from the expression of 70 genes. It has been in the market since 2007 and has been tested on over 3000 patients in the meantime. The original paper underlying MammaPrint® argued that up to 80% of patients receive unnecessary chemotherapy [1] and showed a prediction accuracy of 83%, which was later corrected to 69% [2]. Despite its success, the approach has been criticized. Ein-Dor *et al.* [3] compared the 70 genes signature published by van't Veer *et al.* [1] and a different signature of 76 genes by Wang *et al.* [4]. Ein-Dor *et al.* concluded that 'the overlap between these gene sets is almost zero' and 'many equally predictive lists could have been produced'. Recently, Venet *et al.* [5] continued along these lines and showed in a systematic analysis that 60% of the analyzed signatures

Corresponding author. Michael Schroeder. Biotechnology Center, Technische Universität Dresden, Tatzberg 47-49, 01307 Dresden, Germany. Tel: +49 351 463 400 60; Fax: +49 351 400 61; E-mail: michael.schroeder@biotec.tu-dresden.de

Janine Roy is a PhD student of Bioinformatics at the Informatics Department of Technische Universität Dresden. Her research interest includes network-based analysis methods for cancer outcome prediction.

Christof Winter obtained his PhD in Bioinformatics at Technische Universität Dresden in the group of Michael Schroeder where he was working on structural protein interactions and regulatory networks in pancreatic cancer. Currently, he works as a postdoctoral fellow in the group of Lao Saal at Lund University, Sweden, on breast cancer genomics.

Zerrin Isik obtained her PhD on pathway enrichment systems to assess biological activity of cell processes at Middle East Technical University, Ankara, in 2011. Now she is working in the Bioinformatics group of Michael Schroeder.

Michael Schroeder is the head of the Bioinformatics Department at Technische Universität Dresden. His research interests lie in protein interactions, functional annotation based on text-mining and drug repositioning.

© The Author 2012. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

were not better than random and that more than 90% of random signatures with over 100 genes were significant outcome predictors.

Besides the problem of random signatures, the use of microarrays has been the subject of discussion: Michiels *et al.* [2] showed in seven large cancer prognosis studies that signatures derived by microarray data analysis do not achieve a prediction accuracy better than random and Klebanov and Yakovlev [6] argued that technical noise can become problematic when differences are subtle. On the other hand, a large study of over 50 groups came to the conclusion that technical noise in microarrays is low [7]. Applied to cancer, Dobbin *et al.* [8] concluded that biological variation between tumors exceeds technical variation.

Both problems, i.e. random signatures and noisy expression data, can be addressed by data integration [9]. Assuming that the three independent statements 'A is up', 'A regulates B' and 'B is up' have each an error probability of 10%, then the joint statements have only an error probability of 0.1%. Hence, the error can be reduced by integrating consistent pieces of knowledge. Ideker *et al.* [9] summarized this general principle and also applied it to the above breast cancer data combining gene expression with protein interaction data [10]. This leads to an improvement of 8% compared with the corrected accuracy of van't Veer *et al.* [1].

Several questions arise from the studies: Does network outcome prediction provide consistent improvements across different cancer types? Does the size of the signature and patient cohort influence these results? Which types of interactions—physical or regulatory—are best suited? Are current networks sufficiently large? If the generated signatures are less noisy, are they better than random ones? How do the accuracies compare against each other? In this work, we address these questions by applying a network-based outcome prediction algorithm, NetRank, based on the PageRank algorithm [11] to 25 cancer datasets. The NetRank algorithm has been successfully applied on pancreatic ductal adenocarcinoma to predict survival time of patients [12]. The aim of this study is to investigate and evaluate the outcome prediction performance of NetRank on different cancer datasets.

RELATED WORK

Network information efficiently helps to improve outcome prediction and reduce noise in microarray experiments [13]. For this purpose, network information has been integrated with microarray data in various studies in the last decade. A pioneering approach for integration of network information and microarray data has been developed to define subnetwork activity as the aggregate expression of genes in a given subnetwork [10]. Several other studies have aimed to identify significant subnetworks based on given expression data [14–19].

The PageRank algorithm also uses network information (i.e. hyperlinks) to rank the pages in the World Wide Web according to their importance [11]. Several recent studies have applied the idea of the PageRank algorithm for both classification of cancer tumor data [20–24] and genome-wide association studies [25–27]. The idea of using the PageRank algorithm for microarray analysis was first introduced by Morrison *et al.* [22] and successfully applied by Winter *et al.* [12] on pancreatic ductal adenocarcinoma. The suggested algorithm assigns a rank value to a gene that is a combination of its score and the scores of genes linked to it. Basically, if a gene is linked to other highly ranked genes, it will also get a high rank due to the boosting effect of neighboring genes. Hence, the algorithm provides an integration of the topological information (i.e. connectivity and random walk) and microarray data (i.e. node score) to explore crucial signature genes for outcome prediction. Another study applies the original PageRank algorithm to explore predictive subnetworks by identifying cross-talking proteins between seed genes obtained from other cancer-related studies [23]. Hai *et al.* [20] and Osmani *et al.* [28] use the PageRank algorithm to identify characteristic genes specific for a disease in a regulatory network. An approach similar to the NetRank algorithm aims to identify significant genes from an interaction network to predict breast cancer risk of patients [21]. A pathway-based approach explores the effects of gene expressions on the pathways and provides an impact score for each pathway under the specific disease conditions [24]. In this study, we use an adapted version of the PageRank algorithm, called NetRank, for cancer outcome prediction [12]. Our aim is to provide a comprehensive evaluation of the NetRank algorithm on several cancer datasets by comparing its performance to the classical methods. Furthermore, we analyze the overlap of signatures between cancer and outcomes as well as the effect of the different types of network (regulation/interaction) on the prediction accuracy.

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

RESULTS

NetRank and data

NetRank: googling for biomarkers

The PageRank algorithm assumes a random surfer who navigates through a network of web pages by following links with probability d or by starting new sessions on a random page with probability $1-d$. The more frequently a page is visited by the surfer, the higher becomes its rank value. The damping factor d influences how strong the impact of the network is, since it balances link navigation (network) and random restarts. In contrast to other algorithms for ranking nodes in networks [16, 29, 30], the parameter d in the PageRank algorithm can be used to assess how important the network information is for the ranking task. Since this is one of the open questions posed above, we use a PageRank variant for outcome prediction as published in Winter *et al.* [12]. In the NetRank algorithm, the damping factor is optimized for each dataset to obtain better outcome predictions. A summary of the NetRank is provided in the Methods section and in Figure 5.

25 cancer datasets

We carried out a substantial literature search to identify suitable datasets for cancer outcome prediction. A search on PubMed with relevant keywords relating to cancer, gene expression and outcome (detailed in 'Methods' section) returned 523 articles. From these, we selected publications in journals with impact factor >5. The remaining 239 articles were screened manually. The most widely used chips were HGU133plus2 and HGU133A. For the sake of comparability, we only retained papers using these chips. This resulted in 25 suitable datasets summarized in Table 1.

The 25 datasets cover 13 different types of cancer including lung, breast, leukemia, lymphoma and prostate cancer. They differ in the outcome variable (diagnosis, response to treatment, progression or subtyping), the size of signatures (from 1 to 357 genes), their prediction accuracy according to the authors (from 60% to 100%) and the methods employed [expression analysis with *t*-test, fold change, Significance Analysis of Microarrays (SAM), Prediction Analysis of Microarrays (PAM) and the use of machine learning and clustering]. None of these studies employed network-based approaches and only 13 studies reported prediction accuracies. However, two of these [40, 41] are not directly comparable since they did not separate training and test data samples, so we considered only 11 studies with reported accuracies.

Accuracy and signature/cohort size

An interesting question that arises directly is whether the size of signature and patient cohort have an influence on the results. Venet *et al.* [5] showed that any random signature has a predictive power if it is large enough. Hence, we assessed whether signature size and prediction accuracy are correlated. Also, a small cohort could be prone to more artifacts and hence have a better accuracy than a large and more representative study. Therefore we explored these issues by using the original accuracies as provided in 11 publications. While investigating the effects of these parameters on the prediction accuracy of the studies, we could not observe such clear correlations: $r=-0.46$ for signature size versus accuracy and $r=-0.14$ for cohort size versus accuracy. As can be observed in Figure 1A, prediction accuracies are not related to cohort size. Bogunovic *et al.* [36] achieve with 38 patients a similar accuracy as Steidl *et al.* [34] with 130 patients. It also has been shown that an exact outcome prediction is possible with a small cohort of only 27 patients [40]. The same can be considered when comparing prediction accuracy with signature size (Figure 1B). A large signature does not ensure a high prediction accuracy.

Next, we discuss how the network-based approach compares against the classical approaches and how the underlying network affects the results.

NetRank and other methods

NetRank is better than random and classical approaches

Table 1 provides the accuracy comparison of two classical approaches (fold change and *t*-test) with NetRank based on two different types of networks. Additionally, the sensitivity and specificity values are provided in Supplementary Table S1. NetRank always performs better than random leading to an average improvement of 12% (ranging from nearly 0% to over 40%). In nearly all cases (23/25), NetRank is also better than the classical methods. The average improvement here is 6% ranging between 0% and 13%. The notion of network information improving outcome prediction has been shown before. However, these studies focused mainly on a limited number of cancer types (e.g. breast [10, 21], breast and colon [15]). Here, we show that a network-based outcome prediction

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

Table 1: Results for 25 datasets covering 13 cancers, four outcome tasks and different cohort sizes

Author	Year	Cancer	Var.	#Pat.	Acc[a]	Non-network		Network-based		Random	Accuracy Improvement			Best
						FC	*t*-test	HPRD	Transfac		net/no-net	net/orig	net/random	d-value
Jones *et al.* [31]	2005	Renal cell	S	69[b]	100	96	96	97	97	79	1	-3	18	0.1
Spira *et al.* [32]	2007	Lung	D	129	82	66	67	72	69	65	4	-12	6	0.8
Lee *et al.* [33]	2010	Bladder	P	165	79	70	69	73	76	56	5	-5	19	0.1
Steidl *et al.* [34]	2010	Lymphoma	T	130	77	72	75	73	74	70	0	-4	4	0.1
Bhojwani *et al.* [35]	2008	Leukemia	T	82	74	64	58	63	64	51	3	-11	13	0.1
			P	59	73	67	64	70	72	54	6	-2	17	0.4
Bogunovic *et al.* [36]	2009	Melanoma	P	38	74	53	63	64	70	56	9	-7	11	0.9
Raponi *et al.* [37][c]	2006	Lung	P	129	64	59	55	64	59	54	5	-3	8	0.2
van de Vijver *et al.* [38][c]	2002	Breast	P	295	62	61	58	66	68	62	8	5	5	0.7
Wang *et al.* [4]	2005	Breast	P	276[b]	62	66	65	70	67	66	3	7	3	0.5
Korkola *et al.* [39]	2009	Germ cell	P	82[b]	60	67	64	70	71	67	5	11	5	0.2
O'Donnell *et al.* [40]	2005	Oral cavity	P	27[b]	100[d]	66	78	81	81	75	9	-	6	0.2
Friedman *et al.* [41][c]	2009	Leukemia	P	68	65[d]	52	60	57	55	53	0	-	3	0.1
Landemaine *et al.* [42][c]	2008	Breast	P	23	-	95	90	100	99	67	7	-	33	0.3
Nanni *et al.* [43][c]	2006	Prostate	D	30	-	89	96	100	99	90	7	-	10	0.5
			T	20	-	55	62	72	70	31	13	-	40	0.8
Iqbal *et al.* [44]	2010	Lymphoma	D	80[b]	-	76	77	81	82	67	5	-	15	0.2
Fernandez *et al.* [45]	2010	Lymphoma	S	22[b]	-	78	67	80	81	44	8	-	37	0.2
Frank *et al.* [46]	2006	Leukemia	T	41[b]	-	70	67	70	75	66	4	-	7	0.8
Smith *et al.* [47]	2010	Colon	P	55	-	68	57	72	71	62	9	-	10	0.2
Lenz *et al.* [48]	2008	Lymphoma	T	414	-	60	59	63	65	60	5	-	4	0.3
Mok *et al.* [49]	2009	Ovarian	P	53	-	60	54	64	65	65	8	-	0	0.1
Dressman *et al.* [50]	2006	Breast	S	37	-	51	46	55	64	43	11	-	17	0.9
Murat *et al.* [51]	2008	Glioblastoma	T	70[b]	-	49	46	58	59	48	11	-	11	0.3
Zhu *et al.* [52]	2010	Lung	P	133	-	50	46	53	58	50	8	-	6	0.7

Accuracy %: 30-34 | 35-40 | 31-45 | 46-50 | 51-55 | 56-60 | 61-65 | 66-70 | 71-75 | 76-80 | 81-85 | 86-90 | 91-95 | 96-100

The comparison shows consistent accuracy improvement of NetRank over random (column *net**random*) signatures and classical methods (*net**no-net*).

[a]Accuracy of original studies. If it was not provided by theauthors and if applicable, it was calculated by Equation 2.

[b]A subset of the original dataset was used for NetRank validation.

[c]These datasets did not provide raw data.

[d]No separation between training and test patients.

clinical variables: D - Diagnosis, P - Prognosis/Progression, S - Subtyping, T - Treatment Response / Outcome

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

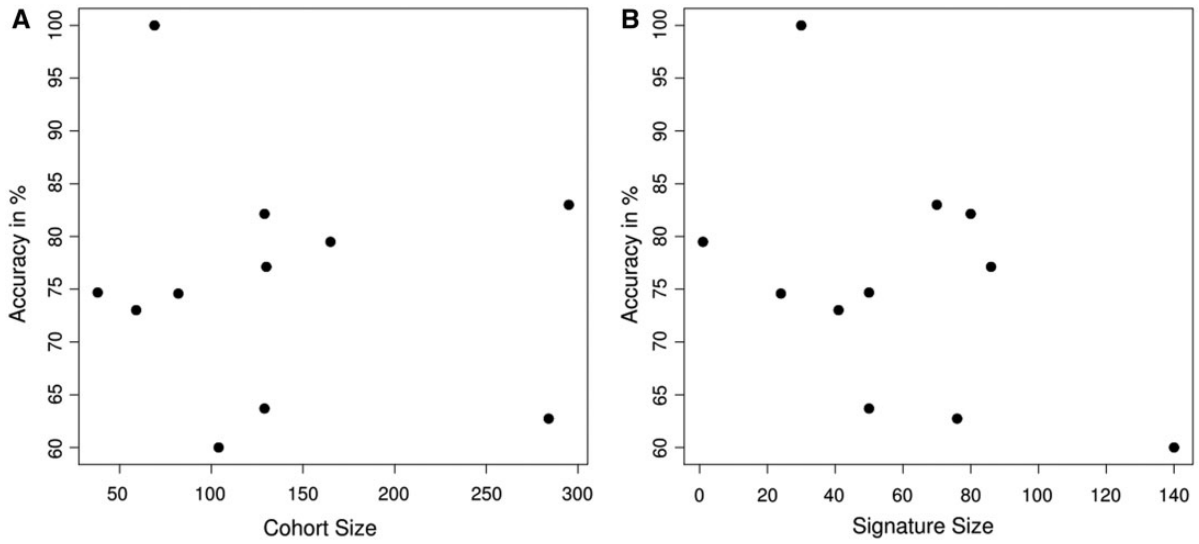

Figure 1: Accuracy vs. cohort size (**A**) and signature size (**B**). Prediction accuracy does not correlate with the number of patients (correlation −0.14) nor with amount of genes in signature (correlation −0.46). The analysis is based on 11 accuracies, which were provided in the original publications.

algorithm is able to predict outcome in many different types of cancer. In addition, our result confirm that network information improves outcome prediction in comparison to classical methods like *t*-test and fold change.

NetRank is on a par with authors' results

None of the authors of the 25 datasets employed network-based approaches and only 13 studies reported prediction accuracies. We considered only 11 studies with reported accuracies due to no separate training and test samples in two studies [40, 41]. How does NetRank compare to these studies? Caution must be taken here. First, the authors' own computation of accuracy is not directly comparable to our definition (see 'Methods' section). Second, most of the reported accuracies are substantially better than classical methods (9/11 are better; 9% average improvement). This means that most of the authors apply advanced analysis methods in the selection of signature genes. For example, Steidl *et al.* [34] applied a sparse multinomial logistic regression to construct signatures in combination with clinical variables. Bogunovic *et al.* [36] derived a predictive signature by using principal component analysis. Lee *et al.* [33] assigned E2F1 as the predictor gene to identify the invasive progression of bladder tumors by using up-modulated gene expression and prior knowledge. Nonetheless, we compared NetRank to the author accuracies: For 3 out of 11 NetRank is better, but on average NetRank is 2% worse. Taken together, this means that NetRank's improvement over classical methods is approaching the level of the authors' signatures.

Physical interactions as good as regulation

For the above results, we tested two networks: Human Protein Reference Database (HPRD) [53] and Transfac [54]. They differ in their type and size. HPRD covers physical interactions and Transfac is a network of transcription factors and targets. HPRD contains nearly 10 000 genes and 40 000 interactions, whereas Transfac has only some 2400 genes and 5700 interactions. As shown in Table 1, there is no substantial difference between these networks (Transfac is better than HPRD in 16/25 cases, average improvement 1%). Nevertheless, employing only the Transfac network achieves an equal performance using far less genes. Cancer often arises due to alterations in transcription factor expression, leading to unregulated cell growth and differentiation. Therefore, the efficiency of Transfac in cancer outcome prediction is biologically reasonable. This might implicate that future regulatory networks lead to further improvement in outcome prediction.

Prognosis and treatment is more difficult than diagnosis and subtyping

The main goal of outcome prediction lies in defining the future state of a patient based on its current disease state. Outcome prediction mainly focuses on the

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

following topics: (1) Treatment response deals with the different response types of patients to treatments. Response means a prolonged metastasis-free survival or faster response to therapy. (2) Prognosis/Progression deals with cancer development after (3) Diagnosis. Finally, (4) Subtyping copes with finding subtypes of cancer. Subtypes may alter in the aggressiveness of the cancer along with different spreading behavior and may end up in different treatment needs. Figure 2A summarizes NetRank's average accuracies (Table 1) for different outcome prediction tasks. As the figure shows, diagnosis and subtyping are easier than response to treatment and prognosis. One explanation for this result could be that treatment and prognosis depend more on external, non-molecular parameters such as age, sex, living conditions, etc. Another possible explanation is the difference in gene expression between outcome groups. When diagnosing a malignant disease, healthy and cancerous tissues are compared. These two types of tissue express highly different number of proteins [55], whereas in the case of progression and treatment response prediction, the difference in the gene expressions is very subtle.

Network influence

As discussed in the Introduction section, one objective was to assess the influence of the network on the outcome prediction accuracy. In NetRank's random surfer model, the damping factor d, which is optimized for the individual prediction task, allows to judge the network influence. With a $d=0$, the surfer never follows links and hence the network topology has no influence, whereas a $d=1$ implies a surfer who only follows links without randomly restarting and thus increases the importance of the network. The best damping factor for each dataset is computed via leave-one-out cross-validation values from 0 to 1 in steps of 0.1. Figure 2B shows the accuracy improvement of NetRank over the classical approaches plotted against the best damping factor. The figure shows a bimodal distribution. For 12 of 25 networks, d is <0.3 and for 5/25 networks d is >0.7. This implies that the underlying networks are not of equal importance in all tasks. One reason could very simply be that network information is still sparse. HPRD's 40 000 interactions and Transfac's 5700 represent only a small fraction of the estimated 650 000 interactions of the human interactome [56]. If the coverage of the network is a key factor, then one would expect a correlation between the damping factor and the accuracy improvement. This is indeed the case with a correlation of $r=0.62$, meaning by improving network coverage, outcome prediction results will be increased. Besides, Transfac and HPRD might include spurious interactions due to the errors in high-throughput experiments, however such errors are less important than false interactions predicted by text-mining methods [57, 58]. Under the assumption of independence of experimental errors in expression and interaction data, the outcome prediction is clearly improved by using network information.

Influence of cancer bias

Network information is generally extracted from literature studies. Therefore genes in such networks are often highly related to certain diseases, e.g. cancer. In order to assess the bias of well-studied proteins in a network, we decided to limit the analysis to these genes. For this purpose, we solely provided genes mentioned in Transfac as an input for the classical methods (fold change, t-test) and assessed their prediction accuracy via leave-one-out cross-validation. As indicated in Supplementary Table S2, the prediction accuracies of the classical methods obtained by this limited gene set are generally lower than NetRank results based on Transfac. The results indicate that the information gain achieved by NetRank is not solely based on well-studied cancer proteins but also on the interaction between them. Hence, it provides more predictive signatures compared with the classical methods.

Influence of the machine learning method

During experiments, we were faced with the problem of prediction accuracies significantly lower than 50%. For instance, Support Vector Machines (SVMs) could not clearly be trained on the Mok *et al.* and Nanni *et al.* datasets [43, 49]. Notably, the few patient samples ($n=20$) in the treatment dataset of Nanni *et al.* might be a reason of the poor performance of SVM. Therefore, we assumed that other machine learning approaches would be more suitable. To assess the prediction performance of different machine learning approaches, we trained them using the signatures obtained by NetRank in one dataset and tested them on all other datasets (Figure 5: Steps 5–7). The results for SVM, linear discriminant analysis (LDA) and k-nearest neighbor (k-NN) are given in Figure 3. The performance of Random Forest (RF) was almost similar to LDA and decision tree (DT) delivered low prediction accuracies (results

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

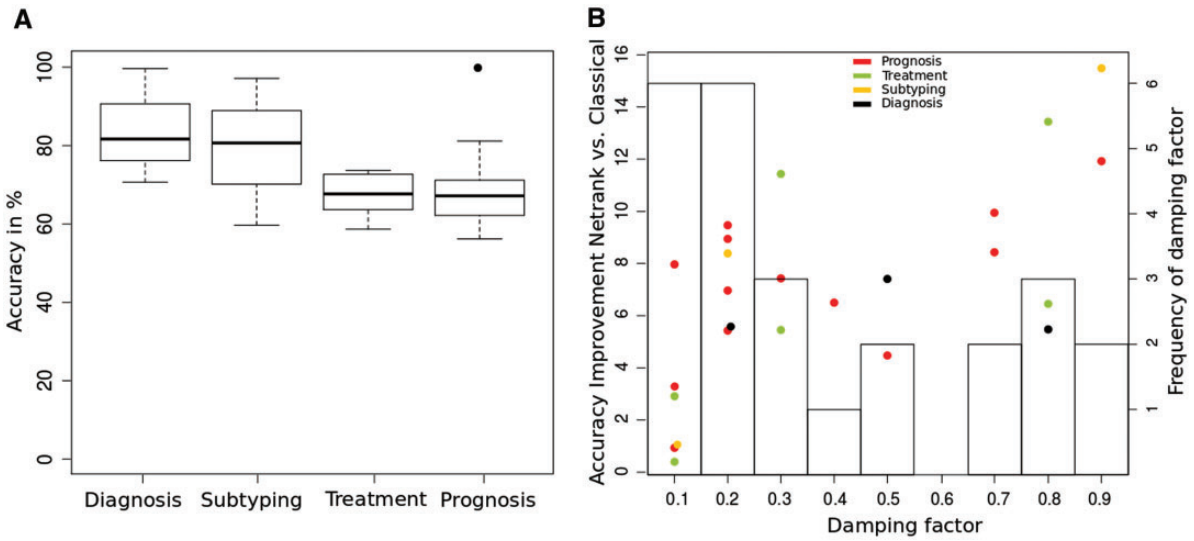

Figure 2: NetRank accuracy vs. outcome task (**A**) and damping factor (**B**). Diagnosis and subtyping are easier than treatment or prognosis (A). The damping factor measures network influence. The factor is optimal for each dataset. Overall, stronger network influence leads to better accuracy improvements (the correlation is 0.62). Note: The average prediction accuracies of datasets (Table 1) were taken as the base in these plots.

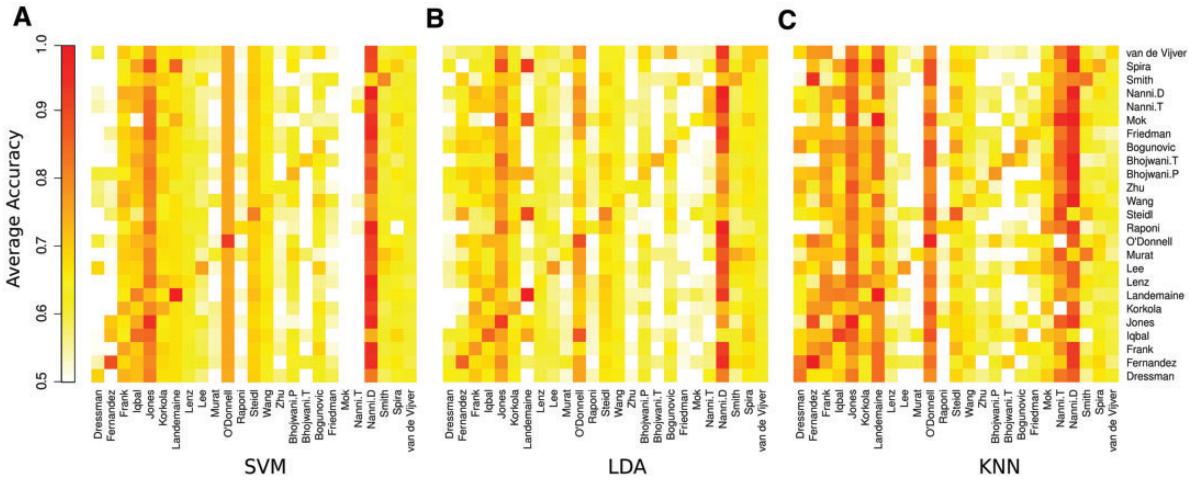

Figure 3: Accuracy heatmap of transferred signatures with different machine learning methods. Signatures were found with NetRank on Transfac. Rows and columns represent the predictive signatures and datasets, respectively. So, the diagonal shows the accuracy of the signature on the original dataset. The accuracies lower than 50% (worse than a random guess) are indicated in white color. Different machine learning techniques gave different accuracies. SVMs have an overall lower accuracy and fail to predict progression outcome of Mok *et al.* and Nanni *et al.* datasets [43, 49]. For our datasets, the k-NN approach has an overall higher accuracy than SVM and LDA.

not shown). For our datasets, the k-NN approach has an overall higher accuracy than SVM and LDA (3% and 2% average improvement). We did not replicate the analysis with the k-NN learning method, due to the computational intensity of each experiment. These experiments show that the prediction performance of the NetRank would be further improved by using several machine learning methods. However, as stated in the *No-free lunch theorem* [59], there is no ideal machine learning method providing a high classification accuracy independent of context and data.

Reproducibility of signatures

Ein-Dor *et al.* [3] criticized that the signatures for breast cancer outcome prediction published in van't

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

Veer *et al.* [1] and Wang *et al.* [4] have an overlap of almost zero and that other 70 gene signatures work with a similar accuracy. Is this situation similar regarding signatures reported by authors and signatures computed by NetRank?

To address these questions, we plotted the pairwise similarity of signatures (see 'Methods' section for similarity definition) against each other. In Figure 4, a value of 0 (white) indicates no overlap and 1 a strong overlap between the signatures. We also computed *P*-values of Jaccard indices using Fisher's exact test (see 'Methods' section), which shows the probability of an overlap between the signatures by chance. The −log(*P*-value) of each pairwise overlap is plotted in Supplementary Figure S1.

Author signatures do not overlap

As Figure 4A shows, there is no overlap between the signatures published by the authors. Exceptions are O'Donnell *et al.* [40] and Nanni *et al.* [43], which share seven transcription factors (out of 205 genes). The lack of overlap applies also to signatures for the same cancer type supporting the finding of Ein-Dor *et al*. For *t*-test and fold change as classical methods, a similar conclusion can be drawn (data not shown).

Network signatures do overlap

In contrast to author signatures, network signatures have a significant overlap (Figure 4B and C). This indicates that NetRank selects similar genes for the prediction independent of the cancer type. As the figure shows, this effect is even stronger for the Transfac network, which can be explained by the smaller size of Transfac compared with HPRD. Interestingly, all signatures share on average 3.2 genes with each other. In the case of the Transfac network, the most commonly selected genes are SP1, FOS and JUN. These findings correlate with a recently published study on pancreatic cancer [12]. It has been shown that these genes are related to cancer development and progression. For example, FOS and JUN oncogenes form the AP-1 complex that controls several cancer-critical processes such as proliferation, migration and survival [60]. Several studies discussed the crucial role of SP1 in different cancer types, e.g. breast, pancreas, colon and prostate cancer [61–64].

In the same vein, other studies discovered a high overlap between predictive signatures. Dutkoskwy and Ideker [65] developed a DT-based approach by integrating expression and interaction data to identify dysregulated subnetworks. They found a high correlation (73%) between the predictive genes of two breast cancer datasets and claimed that predictive signatures would reveal underlying mechanisms of the diseases. Therefore as a future work, we will explore the key roles of such similar signatures in the cancer progression.

While most signatures are strongly overlapping, Raponi *et al.* [37] is different from all of them (Figure 4C). One reason could be that Raponi *et al.* studied a combination of squamous and adenocarcinoma non-small-cell lung cancer samples. Such a heterogeneity might also be the reason why other signatures achieve consistently poor predictions on this dataset (Figure 3A and C).

Comparison with other network-based methods

NetRank provided better outcome predictions compared to non-network-based methods, hence it is an interesting question whether other network-based methods also lead to improvements. To test this question, we compared the prediction performance of NetRank with other network-based methods [10, 66] on two breast cancer metastasis datasets [4, 38] (Supplementary Table S4). In this comparison, NetRank accuracies based on Transfac are used and the accuracies from the two network-based methods are taken from the respective publications [10, 66].

Chuang *et al.* [10] integrated network information and cancer data by defining subnetwork activity as the aggregate expression of genes in a given subnetwork and used greedy search to identify subnetworks with discriminative power. Chen *et al.* [66] extracted subnetworks to classify metastatic patients by combining network-constrained support vector machines with significance test based on label permutations. The prediction results of the three network-based methods on two breast cancer datasets is shown in Supplementary Table S4. Variances in the sensitivity and specificity measures of the three network-based methods might be originated by the usage of different machine learning methods and cross-validation procedures during the training of the classifiers. These results confirm that network information improves outcome prediction in comparison to the classical methods as well as to the original results [4, 38].

DISCUSSION AND CONCLUSION

Although the first diagnostic kits for cancer outcome prediction are in the market allowing patients to

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

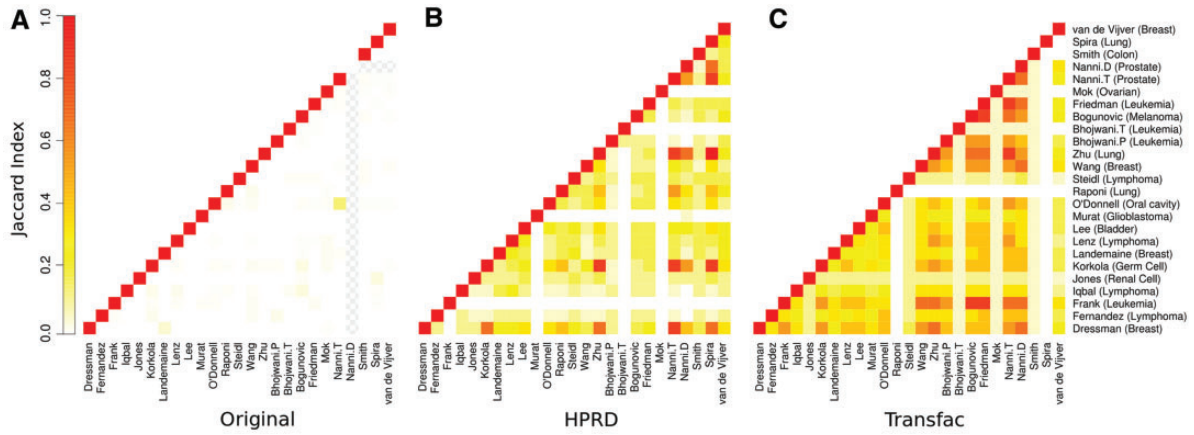

Figure 4: Pairwise comparison of signatures. (**A**) Signatures of authors do not overlap—even for the same cancer. (**B**) Signature of NetRank with physical interaction and (**C**) regulatory network strongly overlap. *Note:* No signature was published for diagnosis in Nanni *et al.* [43].

personalize treatment options, the criticism remains that large signatures appear random, since other signatures perform equally well. Here, we aimed to address this problem by focusing on signatures that are compact (only 10 genes) and that were obtained with a network-based approach, which puts the signature genes into a context rather than selecting isolated genes. Indeed, we found on 25 datasets covering different cancers, outcome tasks and cohort sizes that there is a consistent and significant improvement of the network-based approach over random signatures and a classical approach based on *t*-test or fold change. Besides, we are approaching the accuracy level of the authors' signatures by applying a relatively unbiased but fully automated process for biomarker discovery. We discovered that predicting response to treatment and progression are generally more difficult than diagnosis and subtyping. We hypothesized that the former is more strongly influenced by external factors such as age and sex. Furthermore, we conclude that in diagnosis the gene expression signal is much stronger. This can be proven by comparing the random accuracy of Nanni *et al.* [43] with any other accuracy obtained by more sophisticated methods. The signal is strong enough to be captured by random gene selection.

An important open question is also the influence of the network topology on the analysis. We experimented with a regulatory (Transfac) and physical interaction (HPRD) network. The latter is an order of magnitude larger than the former. Nonetheless, both can be considered as a fraction of the complete interactome that is currently not yet known. Interestingly, despite the size difference, they perform equally well. This suggests that regulatory information is particularly suited for outcome prediction. Most cancer types arise due to the mutations in regulatory elements, therefore the efficiency of Transfac in cancer outcome prediction is biologically reasonable.

The influence of the network is also visible in the choice of optimal damping factor *d*, which balances expression and interaction. A weak correlation of *d* to accuracy improvement was observed, suggesting that increasing coverage of the interactome may also lead to further improvements in the prediction accuracy. Furthermore, we evaluated the disease bias that genes in the networks are often highly related to certain diseases and we showed that this bias does not affect the prediction abilities of NetRank. Finally, we investigated the influence of machine learning approaches on outcome prediction. We could clearly observe a difference between diverse machine learning approaches. Overall, the k-NN approach performs best, but it was computationally infeasible to fully test the effect of k-NN on our results.

As a summary, network-based gene expression analysis is leading to a more detailed understanding of cancer and cancer-related processes by selecting highly relevant genes that are not just correlating with but actively influencing the outcome of a patient. Furthermore, putting prognostic signatures into the context of pathways and network neighborhood may provide crucial information to move from biomarkers to targets, whose modulation will influence outcome.

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

METHODS

Outcome datasets

We ran a query (Supplementary Table S3) in PubMed on 1 April 2011. We only retained papers published in journals with an impact factor >5 [by Thomson Reuters (ISI) Web of Knowledge, version 2009] and datasets using Affymetrix HGU133plus2 or HGU133A chips. If raw microarray data were available, we used the RMA implementation as provided in the *affy* R package, which automatically performs background correction and summarization and then transforms the expression values into $\log_2$ scale.

Genes and nodes in networks are represented by Entrez gene identifiers, therefore probes with unknown Entrez identifiers were discarded. If a gene is represented by several probesets, the probeset with the highest mean expression was selected as the representative for that gene [67].

Networks

We used human transcription factor–target pairs as provided in the Transfac Suite 2011 [54] as well as the HPRD [53]. Transfac contains manually curated regulations between transcription factors and targets. HPRD covers physical protein–protein interactions manually compiled from experimental studies in the literature.

Development of prognostic signatures via Monte Carlo cross-validation

In order to predict signatures, we applied a Monte Carlo cross-validation workflow (Figure 5) that evaluates the accuracy of classification methods to rank genes for outcome prediction. In this procedure, the dataset is randomly divided into a training and a test set (Step 2). Within the training set, NetRank ranks genes by how different they are between patients' outcome classes (Step 3). The 10 highest ranked genes are selected as signature and used to train a machine learning classifier (Steps 4 and 5). Genes in signatures were always represented as Entrez gene identifiers and the comparisons are made by Entrez identifiers. Afterwards, the classifier is used to predict the outcome of the test set patients (Step 6). The accuracy is computed by comparing predicted outcome with the true outcome (Step 7). This workflow (Steps 2–7) is repeated 500 times with new random splits of data and the average accuracy is provided as outcome prediction performance of the method. This workflow is a recommended and relatively unbiased evaluation strategy [2].

We set the signature size to 10 genes for an easy experimental validation. One approach to use biomarkers in the clinical routine is the application of immunohistochemistry as performed in our previous study [12]. Such antibody staining is only feasible for a small set of genes. Additionally, the fixed signature size provides a better comparability of different datasets. A cross-validation approach to find the optimum signature size for each dataset would not be a suitable procedure to explore the real influence of network information on the prediction accuracy. Various feature selection studies for microarray data also showed that training of a classifier with larger size of signature genes did not provide a significant improvement on the prediction accuracy [68, 69].

Poor and good prognosis

Patients in the dataset were divided into two groups based on their clinical outcome. If the outcome class of patients was not explicitly defined in the dataset, patients were classified based on median survival time. In that sense, poor prognosis reflects a survival or response time less than a certain time threshold (e.g. 3 or 5 years), whereas patients who survive longer than the median survival time are classified into the good prognosis group.

Classification procedures

A prognostic model with 10 signature genes was developed by employing a support vector machine. The method classifies the tumor samples into poor or good prognosis groups based on the expression levels of the selected genes. Here, we used the implementation as provided in the R package e1071 (version 1.5-18, obtained July 2010). The expression of each gene was used as an independent feature to train the classifier. No aggregation was applied. All feature selection and machine learning steps were subjected to Monte Carlo cross-validation.

NetRank

NetRank is based on the PageRank algorithm [11]. It assumes a random surfer following links or starting a new session. As the surfer navigates through the network, he propagates the value of a gene to his neighbors. Upon starting a new session, a gene is initialized using its *t*-statistic or the fold change value. After several random walks on the network, the rank value of each gene convergences and the final ranks are provided as output.

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

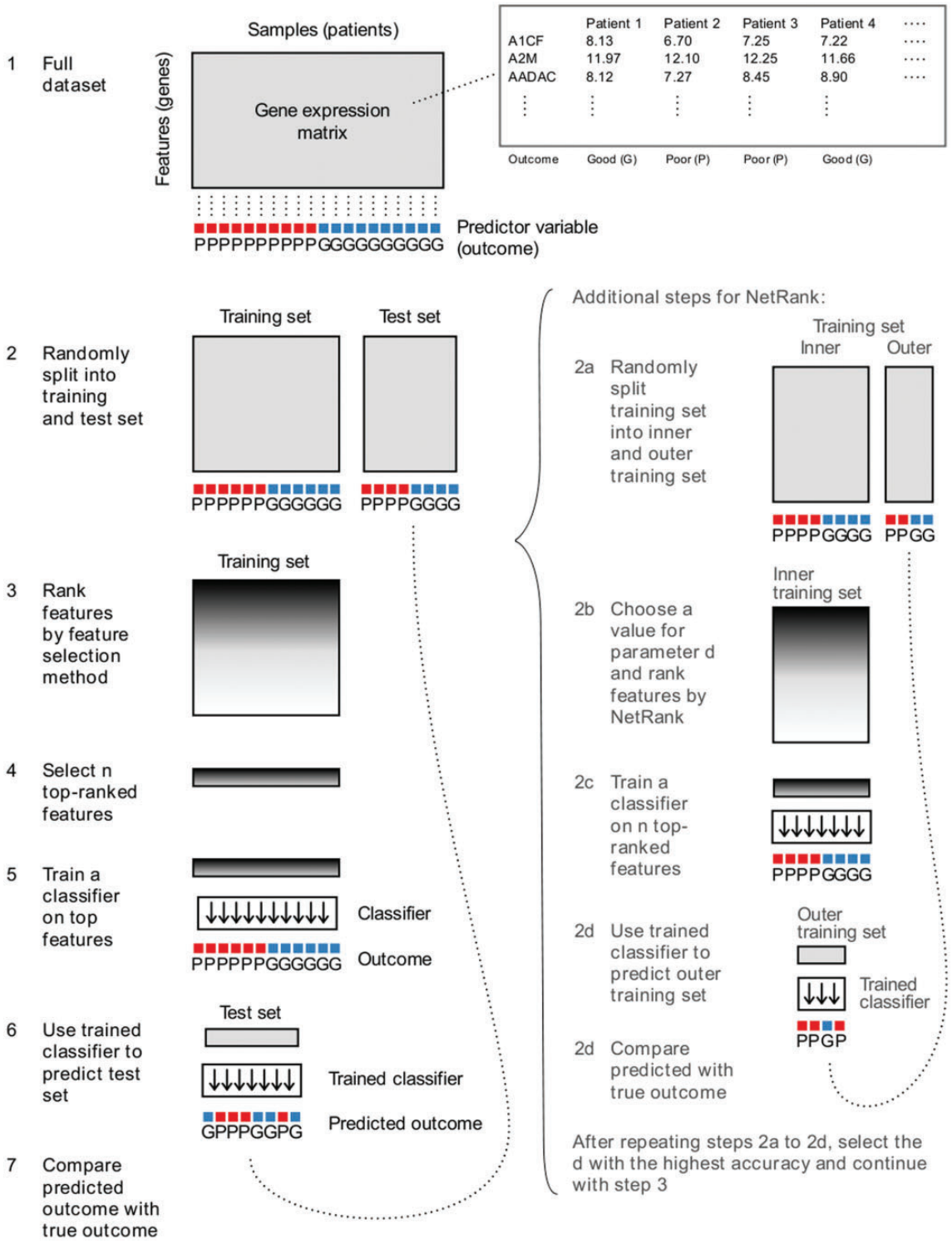

Figure 5: The workflow of NetRank (adapted from Winter *et al.* [12]).

Let us assume that there are N genes in a network. The linkage of these genes, i.e. the network, is stored in an $N \times N$ adjacency matrix M, where $m_{ij} = 1$ if genes i and j are connected and $m_{ij} = 0$ otherwise.

The parameter d is the damping factor, which sets the ratio of starting a new session $(1 - d)$ and following a link (d). When the surfer starts a 'new session' at gene j, the gene is initialized with a network-independent value assessing its importance. Here, this is s_j, which is either the t-statistic value of t-test or the fold change of gene j between different outcome groups. As the surfer follows links, the

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

gene's value is set to the average value of its neighbors from the previous iteration. Thus, the rank r_j^n of gene j in the nth iteration is defined as follows:

$$r_j^n = (1-d)s_j + d\sum_{i=1}^{N}\frac{m_{ij}\times r_i^{n-1}}{\text{degree}_i}, \quad 1 \le j \le N \tag{1}$$

r_j^n convergences to a unique solution which corresponds to the Eigenvalue decomposition [22]. Note that the parameter d is set as part of the Monte Carlo cross-validation workflow (Step 2a–d). For NetRank, we added an additional inner cross-validation loop. In this inner cross-validation, a part of the training set samples was set aside, and different values of d ranging from 0 to 1 in steps of 0.1 were used to run the NetRank algorithm on the remaining training set samples. Accuracies of the top-ranked genes were then tested on the samples previously set aside. As a result of this inner cross-validation, one d value was chosen and used once for deriving a signature using the whole training set, and then its accuracy was evaluated on the test set. It is important to note that no information of the test set is used for selecting d, so the choice of an optimal value for d in the inner cross-validation does not rely on any prediction accuracy in the test set data. Furthermore, the optimal d value for each cancer dataset is dynamically identified by applying the inner cross-validation step. Therefore, the dynamic setting of the damping factor is the novelty of the NetRank algorithm.

Accuracy

Accuracy is the percentage of correctly classified samples and it is defined as the proportion of true predictions to all samples.

If no prediction accuracy is stated in the original study, we calculated it by using the following equation:

$$\begin{aligned}\text{accuracy} = {} & \text{sensitivity} * \text{prevalence} \\ & + \text{specificity} * (1 - \text{prevalence})\end{aligned} \tag{2}$$

where the prevalence is determined as

$$\text{prevalence} = \text{number of ill persons in that dataset/dataset size} \tag{3}$$

Signature similarity

Signatures s and t are compared by Jaccard index, i.e. the size of the intersection of s and t divided by the size of their union.

$$\text{JI} = \frac{|\text{A} \cap \text{B}|}{|\text{A} \cup \text{B}|}. \tag{4}$$

The P-value of each Jaccard index is calculated by the Fisher's exact test (*fisher.test* in *stats* R package). This test exploits the probability of an overlap between signatures by chance. In order to provide a better visualization of statistically significant Jaccard indices, we convert the P-value into $-\log(P\text{-value})$ in the heatmap plots.

Random signatures

We created 1000 different random signatures for each dataset. These signatures were created by randomly selecting 10 genes out of all genes in each dataset. We then tested the outcome prediction accuracy of random signatures via Monte Carlo cross-validation on each dataset.

Classical methods

We explored the Student's t-test and fold change as classical methods. Both methods were applied as implemented in the *genefilter* R package. Fold change finds the log ratio of the means of gene expression for two outcome classes. The t-statistic measure provided by the Student's t-test shows the difference between two class means by considering expression variance of the classes. In order to select significant genes, we did not apply either a cutoff threshold or multiple correction test. After calculation of these measures, we filtered out low expression and low variance genes from the potential signature set. All signatures and accuracies for the classical methods were obtained by the same Monte Carlo cross-validation workflow as applied in the NetRank.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Cancer outcome prediction is addressing the problem of forecasting disease progression based on gene expression and thus allowing refined treatment options. Despite first diagnostic kits on the market, there is still open space to improve gene expression based outcome prediction.
- Recent studies show that the integration of prior knowledge such as protein–protein interaction networks comprises the power to increase prediction accuracies.
- We investigate, based on 25 datasets, the notion of network-based outcome prediction and its advantage over classical methods such as t-test and fold change.
- The results demonstrate that a fully automated integration of network information and gene expression data leads to more accurate outcome predictions compared with the signatures of previous studies.

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

Acknowledgements
We kindly acknowledge supercomputing support from the ZIH of TU Dresden and funding by EU, DFG, BMWi (Ponte, Hybris, GeneCloud). We also would like to thank Robert Grützmann and Christian Pilarsky for helpful discussions regarding biological questions.

References

1. van't Veer LJ, Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**(6871):530–6.
2. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005;**365**(9458):488–92.
3. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci USA* 2006;**103**(15): 5923–8.
4. Wang Y, Klijn JGM, Zhang Y, *et al*. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;**365**(9460):671–9.
5. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 2011;**7**(10):e1002240.
6. Klebanov L, Yakovlev A. How high is the level of technical noise in microarray data? *Biol Direct* 2007;**2**:9.
7. Shi L, Reid LH, Jones WD, *et al*. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006;**24**(9):1151–61.
8. Dobbin KK, Beer DG, Meyerson M, *et al*. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin Cancer Res* 2005;**11**(2 Pt 1):565–72.
9. Ideker T, Dutkowski J, Hood L. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* 2011; **144**(6):860–3.
10. Chuang HY, Lee E, Liu YT, *et al*. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007;**3**:140.
11. Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. Technical Report, *Stanford Digital Library Technologies Project,* 1998.
12. Winter C, Kristiansen G, Kersting S, *et al*. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* 2012;**8**(5):e1002511.
13. Fortney K, Jurisica I. Integrative computational biology for cancer research. *Human Genet* 2011;**130**:465–81.
14. Chowdhury SA, Koyutürk M. Identification of coordinately dysregulated subnetworks in complex phenotypes. *Pac Symp Biocomput* 2010;133–44.
15. Dao P, Colak R, Salari R, *et al*. Inferring cancer subnetwork markers using density-constrained biclustering. *Bioinformatics* 2010;**26**(18):i625–31.
16. Fortney K, Kotlyar M, Jurisica I. Inferring the functions of longevity genes with modular subnetwork biomarkers of *Caenorhabditis elegans* aging. *Genome Biol* 2010;**11**:R13.
17. Su J, Yoon BJ, Dougherty ER. Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. *BMC Bioinformatics* 2010; **11(Suppl 6)**:S8.
18. Ulitsky I, Shamir R. Detecting pathways transcriptionally correlated with clinical parameters. *Comput Syst Bioinformatics Conf* 2008;**7**:249–58.
19. Dao P, Wang K, Collins C, *et al*. Optimally discriminative subnetwork markers predict response to chemotherapy. In: *19th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2011), 2011*.
20. Hai DT, Lee W, Thuy HQ. A pageranking based method for identifying characteristic genes of a disease. In: *IEEE International Conference on Networking, Sensing and Control (ICNSC 2008), 2008*;1496–9.
21. Johannes M, Brase JC, Fröhlich H, *et al*. Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics* 2010;**26**:2136–44.
22. Morrison JL, Breitling R, Higham DJ, Gilbert DR. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* 2005;**6**:233.
23. Nibbe RK, Koyutürk M, Chance MR. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput Biol* 2010;**6**(1): e1000639.
24. Tarca AL, Draghici S, Khatri P, *et al*. A novel signaling pathway impact analysis. *Bioinformatics* 2009;**25**(1):75–82.
25. Davis NA, Crowe JEJ, Pajewski NM, McKinney BA. Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine. *Genes Immun* 2010;**11**(8):630–6.
26. Lee I, Blom UM, Wang PI, *et al*. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* 2011;1109–21.
27. Nirmala , Ancha B, Donald S, *et al*. A network-based approach to prioritize results from genome-wide association studies. *PLoS One* 2011;**6**(9):e24220.
28. Osmani GM, Rahman SM. Using Google's PageRank algorithm to identify important attributes of genes. In: *Midwest Instruction and Computing Symposium (MICS 2007), 2007*.
29. Chowdhury SA, Nibbe RK, Chance MR, Koyutürk M. Subnetwork state functions define dysregulated subnetworks in cancer. *J Comput Biol* 2011;**18**(3):263–81.
30. Ulitsky I, Krishnamurthy A, Karp RM, Shamir R. DEGAS: de novo discovery of dysregulated pathways in human diseases. *PLoS One* 2010;**5**(10):e13367.
31. Jones J, Otu H, Spentzos D, *et al*. Gene signatures of progression and metastasis in renal cell cancer. *Clin Cancer Res* 2005;**11**(16):5730–9.
32. Spira A, Beane JE, Shah V, *et al*. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med* 2007;**13**(3):361–6.
33. Lee JS, Leem SH, Lee SY, *et al*. Expression signature of E2F1 and its associated genes predict superficial to invasive progression of bladder tumors. *J Clin Oncol* 2010;**28**(16): 2660–7.
34. Steidl C, Lee T, Shah SP, *et al*. Tumor-associated macrophages and survival in classic Hodgkin's lymphoma. *N Engl J Med* 2010;**362**(10):875–85.
35. Bhojwani D, Kang H, Menezes RX, *et al*. Gene expression signatures predictive of early response and outcome in

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021

high-risk childhood acute lymphoblastic leukemia: A Children's Oncology Group Study [corrected]. *J Clin Oncol* 2008;**26**(27):4376–84.

36. Bogunovic D, O'Neill DW, Belitskaya-Levy I, *et al*. Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proc Natl Acad Sci USA* 2009;**106**(48):20429–34.
37. Raponi M, Zhang Y, Yu J, *et al*. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res* 2006;**66**(15):7466–72.
38. van de Vijver MJ, He YD, van't Veer LJ, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;**347**(25):1999–2009.
39. Korkola JE, Houldsworth J, Feldman DR, *et al*. Identification and validation of a gene expression signature that predicts outcome in adult men with germ cell tumors. *J Clin Oncol* 2009;**27**:5240–7.
40. O'Donnell RK, Kupferman M, Wei SJ, *et al*. Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity. *Oncogene* 2005;**24**(7):1244–51.
41. Friedman DR, Weinberg JB, Barry WT, *et al*. A genomic approach to improve prognosis and predict therapeutic response in chronic lymphocytic leukemia. *Clin Cancer Res* 2009;**15**(22):6947–55.
42. Landemaine T, Jackson A, Bellahcène A, *et al*. A six-gene signature predicting breast cancer lung metastasis. *Cancer Res* 2008;**68**(15):6092–9.
43. Nanni S, Priolo C, Grasselli A, *et al*. Epithelial-restricted gene profile of primary cultures from human prostate tumors: a molecular approach to predict clinical behavior of prostate cancer. *Mol Cancer Res* 2006;**4**(2):79–92.
44. Iqbal J, Weisenburger DD, Greiner TC, *et al*. Molecular signatures to improve diagnosis in peripheral T-cell lymphoma and prognostication in angioimmunoblastic T-cell lymphoma. *Blood* 2010;**115**(5):1026–36.
45. Fernàndez V, Salamero O, Espinet B, *et al*. Genomic and gene expression profiling defines indolent forms of mantle cell lymphoma. *Cancer Res* 2010;**70**(4):1408–18.
46. Frank O, Brors B, Fabarius A, *et al*. Gene expression signature of primary imatinib-resistant chronic myeloid leukemia patients. *Leukemia* 2006;**20**(8):1400–7.
47. Smith JJ, Deane NG, Wu F, *et al*. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* 2010; **138**(3):958–68.
48. Lenz G, Wright G, Dave SS, *et al*. Stromal gene signatures in large-B-cell lymphomas. *N Engl J Med* 2008;**359**(22): 2313–23.
49. Mok SC, Bonome T, Vathipadiekal V, *et al*. A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2. *Cancer Cell* 2009;**16**(6):521–32.
50. Dressman HK, Hans C, Bild A, *et al*. Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy. *Clin Cancer Res* 2006;**12**(3 Pt 1):819–26.
51. Murat A, Migliavacca E, Gorlia T, *et al*. Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *J Clin Oncol* 2008; **26**(18):3015–24.
52. Zhu CQ, Ding K, Strumpf D, *et al*. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J Clin Oncol* 2010;**28**(29):4417–24.
53. Prasad TSK, Goel R, Kandasamy K, *et al*. Human Protein Reference Database—2009 update. *Nucleic Acids Res* 2009; **37**(Database issue):D767–72.
54. Matys V, Kel-Margoulis OV, Fricke E, *et al*. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006;**34**(Database issue):D108–10.
55. Srinivas PR, Verma M, Zhao Y, Srivastava S. Proteomics for cancer biomarker discovery. *Clin Chem* 2002;**48**(8): 1160–9.
56. Stumpf MPH, Thorne T, de Silva E, *et al*. Estimating the size of the human interactome. *Proc Natl Acad Sci USA* 2008; **105**(19):6959–64.
57. Winnenburg R, Wächter T, Plake C, *et al*. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief Bioinform* 2008; **9**(6):466–78.
58. Venkatesan K, Rual JF, Vazquez A, *et al*. An empirical framework for binary interactome mapping. *Nat Methods* 2009;**6**(1):83–90.
59. Duda RO, Hart PE, Stork DG. *Pattern Classification*. 2nd edn. New York: John Wiley & Sons, 2001.
60. Lopez-Bergami P, Lau E, Ronai Z. Emerging roles of atf2 and the dynamic ap1 network in cancer. *Nat Rev Cancer* 2010;**10**(1):65–76.
61. Abdelrahim M, Samudio I, Smith R, III, *et al*. Small inhibitory RNA duplexes for Sp1 mRNA block basal and estrogen-induced gene expression and cell cycle progression in MCF-7 breast cancer cells. *J Biol Chem* 2002;**277**(32): 28815–22.
62. Shi Q, Le X, Abbruzzese JL, *et al*. Constitutive Sp1 activity is essential for differential constitutive expression of vascular endothelial growth factor in human pancreatic adenocarcinoma. *Cancer Res* 2001;**61**(10):4143–54.
63. Hosoi Y, Watanabe T, Nakagawa K, *et al*. Up-regulation of DNA-dependent protein kinase activity and Sp1 in colorectal cancer. *Int J Oncol* 2004;**25**(2):461–8.
64. Chintharlapalli S, Papineni S, Ramaiah SK, *et al*. Betulinic acid inhibits prostate cancer growth through inhibition of specificity protein transcription factors. *Cancer Res* 2007; **67**(6):2816–23.
65. Dutkowski J, Ideker T. Protein networks as logic functions in development and cancer. *PLoS Comput Biol* 2011;**7**(9): e1002180.
66. Chen L, Xuan J, Riggins RB, *et al*. Identifying cancer biomarkers by network-constrained support vector machines. *BMC Syst Biol* 2011;**5**:161.
67. Jordan IK, Marino-Ramirez L, Koonin EV. Evolutionary significance of gene expression divergence. *Gene* 2005; **345**(1):119–26.
68. Xing EP, Jordan MI, Karp RM. Feature selection for high-dimensional genomic microarray data. *Proceedings of the Eighteenth International Conference on Machine Learning* 2001. p. 601–8.
69. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 2006;**7**:359.

Downloaded from https://academic.oup.com/bib/article/15/4/612/405403 by Dokuz Eylul University Library (DEU) user on 21 October 2021