Journal of Biomedical Informatics 43 (2010) 945–952

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Gene pathways and subnetworks distinguish between major glioma subtypes and elucidate potential underlying biology

Stefan Wuchty [a], Alice Zhang [a], Jennifer Walling [a], Susie Ahn [a], Aiguo Li [a], Martha Quezado [b], Carl Oberholtzer [b], Jean-Claude Zenklusen [a], Howard A. Fine [a,*]

[a] Neuro-Oncology Branch, National Cancer Institute, National Institutes of Neurological Disorder and Stroke, National Institutes of Health, Bethesda, MD 20894, USA
[b] NCI Center for Biomedical Informatics and Information Technology, National Cancer Institute, National Institutes of Neurological Disorder and Stroke, National Institutes of Health, Bethesda, MD 20894, USA

ARTICLE INFO

Article history:
Received 12 January 2010
Available online 7 September 2010

Keywords:
Classification
Gliomas
Subnetworks
Pathways

ABSTRACT

Molecular diagnostic tools are increasingly being used in an attempt to classify primary human brain tumors more accurately. While methods that are based on the analysis of individual gene expression prove to be useful for diagnostic purposes, they are devoid of biological significance since tumorgenesis is a concerted deregulation of multiple pathways rather than single genes. In a proof of concept, we utilize two large clinical data sets and show that the elucidation of enriched pathways and small differentially expressed sub-networks of protein interactions allow a reliable classification of glioblastomas and oligodendrogliomas. Applying a feature selection method, we observe that an optimized subset of pathways and subnetworks significantly improves the prediction accuracy. By determining the enrichment of altered genes in pathways and subnetworks we show that optimized subsets of genes rarely seem to be a target of genomic alteration. Our results suggest that groups of genes play a decisive role for the phenotype of the underlying tumor samples that can be utilized to reliably distinguish tumor types. In the absence of enrichment of genes that are genomically altered we assume that genetic changes largely exert an indirect rather than direct regulatory influence on a number of tumor-defining regulatory networks.

Published by Elsevier Inc.

1. Introduction

Gliomas represent a heterogeneous family of primary brain tumors and are a significant cause of cancer mortality in the United States [1]. While the glioma cell of origin is currently unknown, it has been proposed that it may be different from one tumor to another, ranging from a dedifferentiated astrocyte or oligodendrocyte to a fully undifferentiated neural stem cell. The potential different cells of origin compounded by the genetic and epigenetic heterogeneity within gliomas undoubtedly contribute to the heterogeneous biologic behavior and variable clinical course seen in patients with these tumors. As a consequence of this glioma heterogeneity, the standard glioma classification systems which are based on histopathological criteria suffer from significant intra-observer variability and are poor predictors of treatment response and patient prognosis. Furthermore, traditional classification schemas give little insight into the biological basis of the underlying neoplastic process, a particular disadvantage given the advent of molecularly targeted anti-tumor therapeutic agents. Without an understanding of the underlying biology that drives the neoplastic phenotype within a given tumor, devising appropriate clinical trials as well as treating patients will be impossible. Thus, there is a significant need for tumor classification systems that reflect and elucidate tumor biology.

With the availability of high-throughput microarray technology [2,3] gene expression profiles of genes in tumors provide a glimpse into the inner workings of a cell, suggesting the opportunity for a quantitative characterization of individual tumor biology and tumor classification [3–7]. Recently, several groups have identified certain gene-based features that are associated with a particular glioma type. Although important steps forward, the findings of these studies have been limited by the use of expression values that were derived from single gene analysis. Even when classifications are constructed in such a fashion, the derived classifiers are often devoid of biological significance. In fact, the genes of interest are selected through extreme expression patterns and tend to be derivatives of distant primary events, not fully evident in the annotation of the selected biomarkers. Moreover, most complex biological processes, such as tumorigenesis, are a product of concerted deregulation of complex pathways rather than of any single gene. Thus, disease classifications that are based on the analysis of the

* Corresponding author at: Neuro-Oncology Branch, National Cancer Institute, National Institutes of Health, 9030 Old Georgetown Rd., Bethesda, MD 20892, USA. Fax: +1 301 480 2246.
E-mail address: hfine@mail.nih.gov (H.A. Fine).

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2010.08.011

aggregated effects of expression regulation along functional units or functional protein pathways will theoretically be more informative and useful. Such a classification schema should allow a better understanding of the biological differences between the different subgroups of tumors by the functional nature of the discovered classifiers. As proof of principle, we have utilized the expression profiles of glioblastomas and oligodendrogliomas and have identified protein pathways that are enriched or depleted in these different, yet common glioma subtypes. We also utilized a large network of pairwise human protein–protein interactions, protein–DNA interactions and phosphorylation events in order to identify differentially expressed subnetworks that allow the largest possible discrimination between the underlying tumor types. Both approaches hold promise that functionally related gene sets may act as reliable classifiers as well as suggest functional networks that may be uniquely operative in these different types of gliomas.

2. Methods

2.1. Patient selection, tissue acquisition and sample description

All tumor specimens used in this study were obtained from patients undergoing medically indicated surgical resection of a presumed or known glioma. Following written consent in accordance with the National Cancer Institute and the collaborating center's Institutional Review Boards surgical procedures were performed at a number of institutions and hospitals around the country. Specifically, we utilized 134 patient samples collected from a retrospective database at Henry-Ford hospital (HF) and 136 samples from the prospective NCI-sponsored Glioma Molecular Diagnostic Initiative (GMDI) which were provided as snap frozen sections and profiled using HG-U133 Plus 2.0 arrays. In the HF data set, 77 are Glioblastoma Multiforms (GBM) and 57 Oligodendrogliomas (Oligo), while the GMDI data set consists of 81 GBMs and 55 Oligos. Pathological diagnosis was determined by the home instructional neuropathologist and reviewed by two neuropathologist at the NIH who were blinded to the original diagnosis from the home institution. Only tumors that met the criteria of having a consensus pathological diagnosis from the NIH neuropathologists were utilized for our analyses.

2.2. RNA extraction and array hybridization

We used approximately 50–80 mg of tissue from each tumor to determine RNA with the Trizol reagent (Invitrogen, Carlsbad, CA) and verify the RNA's quality with the Bioanalyzer System [8] (Agilent Technologies, Palo Alto, CA) using the RNA Pico Chips. Utilizing T7-linked Oligo (dT) primer, we converted 6 μg of total RNA to cDNA with Superscript reverse transcriptase (Invitrogen) and *in vitro* transcribe complementary DNA with the T7 Bioarray High Yield RNA Transcript Labeling Kit (ENZO Diagnostics) to generate biotinylated cRNA. We fragmented and hybridized 20 μg of purified cRNA to the Genechip® Human Genome U133 Plus 2.0 Expression arrays [9] (Affymetrix, Inc., Santa Clara, CA) and processed the arrays following the manufacturer's recommendations using the fluidics station 450 and high-resolution microarray scanner 3000. Finally, we generated initial gene expression analysis data files using Affymetrix GeneChip Operating Software (GCOS) version 1.3.

2.3. Microarray data preprocessing

Utilizing parameters in .rpt files generated by GCOS, we confirmed that all arrays complied with minimal quality control standards. Specifically, we tested if a scaling factor is <5 when the expression values are scaled to a target mean signal intensity of 500. Similarly, we controlled for a signal intensity ratios of the 3′ to 5′ end of the internal control genes of β-actin and GAPDH being <3. Finally, we required that Affymetrix spike control (BioC, BioDN and CreX) were always present with present call rates of >35% for brain tissue. Arrays that passed the minimal quality control were normalized using dChip [10] at the PM and MM probe level. Using the average difference model to compute expression values, we calculated the model-based expression levels with normalized probe level data. Log-transforming expression values, negative average differences which were deemed biologically irrelevant (MM > PM) were set to 0. Gene expression data sets are available through the Rembrandt database (http://rembrandt.nci.nih.gov/).

2.4. Data treatment

We demanded that the normal specimen section came form non-tumor bearing patients and had no signs of tumor cells on microscopic examination. Furthermore, the behavior of the global gene expression profiles must resemble the normal tissue in exploratory data analysis of microarrays using principle component analysis (PCA) and hierarchical clustering (HC). Accounting for weak signal intensities, we removed all probesets with more than 10% of zero log-transformed expression values. Representing each gene, we chose the corresponding probeset with the highest mean intensity in GBMs and Oligodendrogliomas, separately.

2.5. Genomic alterations

We collected 1979 genes with experimental evidence of at least one genomic alteration. In particular, we utilized a large-scale genomic study of 178 gliomas, allowing the identification of genomic regions affected by copy number alterations (amplifications, homozygous and heterozygous deletions) and allelic imbalances (loss of heterozygosity and gene conversions) [11]. As indicated, mRNA expression and genomic alterations showed strong correlations in many cases, suggesting that a gene's expression is evidently affected by a genomic change. Such candidate genes have been verified using real-time PCR and methylation sequencing assays [12]. As a reliable sequence resource, we utilized the results of a recent re-sequencing effort to identify types of somatic mutations in protein kinase genes in gliomas [13].

2.6. Protein interactions and pathways

We utilized interaction data from large-scale high-throughput screens [14–16] and several curated interaction databases [17–20], totaling 93,178 interactions among 11,691 genes. As a reliable source of experimentally confirmed protein–DNA interactions, we used 6669 interactions between 2822 transcription factors and structural genes from the TRED database [21]. As for phosphorylation events between kinases and other proteins we found 5462 interactions between 1707 human proteins utilizing networKIN [22,23] and phosphoELM database [24]. Pooling all these interactions we obtained a total network of 11,969 human proteins that are embedded in 103,966 links.

As a comprehensive collection of human gene pathways we utilized pathway information from the NCI/NIH/Nature Pathway Interaction Database [25]. Combining data from various sources PID provided information about 1004 different human pathways.

2.7. Detection of significant subnetworks

Utilizing a large network, composed of protein–protein, protein–DNA interactions and phosphorylation events, we applied a greedy algorithm to identify subnetworks that were differentially

expressed in two classes of gene expression profiles [26] as implemented in the PinnacleZ-plugin of the cytoscape package [27]. Specifically, gene expression profiles of samples of two different types (i.e. Glioblastoma Multiforms and Oligodendrogliomas) were transformed into a subnetwork activity matrix. For a given subnetwork M_k, its activity score in sample j, A_{kj} was defined as $A_{kj} = \sum_i Z_{ij}/\sqrt{n}$. While n is the number of genes in M_k, Z_{ij} is the Z-transformed expression score of gene i in sample j that has been normalized over all samples. A greedy search algorithm identified differentially expressed subnetworks if their t-test score maximizes the association between the subnetworks activity score and the two sample classes. Significant subnetworks ($P < 0.05$) were picked according to null distributions obtained by randomizing the topology of the underlying interaction network.

2.8. Gene set enrichment analysis (GSEA)

Utilizing gene sets from the PID database, GSEA [28] allowed us to determine the degree to which genes in a pathway or gene set were largely found at the extreme ends of a ranked list of gene expression, comparing two different tumor samples. Ranking genes according to their t-test scores an enrichment score (ES) was calculated by walking down this ranked list where a running-sum statistic is increased when we found a gene in the underlying gene set and *vice versa*. The enrichment score is defined as the maximum deviation from zero, corresponding to a Kolmogorov–Smirnov-like statistic. The statistical significance of the enrichment score was assessed by a genotype-based permutation test, and we considered a pathway or gene set as significantly enriched if the nominal $P < 0.05$.

2.9. Support-vector machine algorithm (SVM)

SVMs are machine-learning approaches that are widely used for data classification. Each SVM was trained with a set of instance label pairs $(\boldsymbol{X}_i, c_i)$, $i = 1,\ldots,l$ where vector $\boldsymbol{X}$ holds activity coefficients A_{kj} of enriched pathways or differentially expressed subnetworks k in a given sample j that either was assigned to Glioblastoma Multiforms ($c_i = 1$) or Oligodendrogliomas ($c_i = -1$). In our case, training vectors $\boldsymbol{X}_i$ were mapped into a higher dimensional space by a radial kernel function, $k(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp(-g|\boldsymbol{X}_i - \boldsymbol{X}_j|^2)$, where $g = 1$. The SVM algorithm finds a separating hyperplane with the maximal margin in this higher dimensional space, where we set the penalty parameter of the error term $C = 1$. As a fast implementation of SVMs, we used the libsvm 2.8 package [29].

2.10. Significance of links between pathways

Determining the significance of links between pathways if they share proteins that have evidence of genomic alterations, we formed a 2×2 contingency table for each pair of pathways. In particular, we defined α as the number of shared, altered proteins, while β was the number of remaining proteins in a pair of overlapping pathways. Analogously, we defined γ as the number of altered proteins, and δ as the number of remaining proteins in all other pathways. The probability of obtaining any such set of values randomly is given by the hypergeometric distribution

$$p^* = \binom{\alpha+\beta}{\alpha}\binom{\gamma+\delta}{\gamma}\Big/\binom{N}{\alpha+\gamma}$$

where $N = \alpha + \beta + \gamma + \delta$. In order to investigate the two tails of the underlying distribution we constructed all possible contingency tables by keeping the sum of rows and columns constant. The P-value to reject the null hypothesis which is the independence of rows and columns in the contingency table was defined as the sum of the probabilities p_i, of all contingency tables, $P = \sum_{p_i \le p^*} p_i$ [30].

2.11. Feature selection

For each of N tumor samples that are represented by a vector of subnetwork or pathway scores $\boldsymbol{X}$, the I-Relief algorithm [31,32] defines sets of nearest hits H_n (samples of the same tumor type) and nearest misses M_n (samples of the other type). The objective function of the algorithm is to scale each feature such that the average margin in a weighted feature space is maximized. Briefly, the I-Relief algorithm estimates probability distributions of the unobserved data as exponential functions $f(d) = e^{-d/s}$ where we set $s = 2$. Iteratively, I-Relief adopts a quasi Expectation–Maximization strategy to assess the weights of the underlying features until convergence is reached. For details of the theoretical and technical aspects of I-Relief please see [31]. As a reliable implementation of I-Relief we used the mlpy package (http://mlpy.fbk.eu).

In order to determine an optimal subset of features that allows the best possible discrimination between classes, we ranked all features according to their weights. Running through this ranked list, we added a feature to a list of best features and determined their prediction performance. We defined the subset of ranked features with the highest prediction accuracy as our optimized features list.

3. Results

3.1. Classification with enriched pathways and differentially expressed subnetworks

In order to provide proof of principle that a gene set or network-based tumor classification is feasible, we chose to compare two glioma subtypes, Glioblastoma Multiforms (GBM) and Oligodendrogliomas, that are deemed biologically different based on both standard pathologic criteria and clinical course. Utilizing gene expression data from 134 patient samples from Henry-Ford hospital (HF) and 136 samples from the Glioma Molecular Diagnostic Initiative (GMDI) we were interested in finding groups of genes that allow a reliable discrimination between the underlying tumor types. Although traditional gene expression-based classification approaches have evaluated individual genes, one could potentially use groups of genes to identify pathways that are over- or under-expressed in one tumor type compared to the other (Fig. 1a). Utilizing 1004 fully annotated pathways from the Pathway Interaction Database (PID), we applied Gene Set Enrichment Analysis (GSEA) [28] to determine pathway enrichment in GBMs compared to Oligodendrogliomas and identified 196 pathways in the HF data set and 122 in the GMDI data set ($P < 0.05$, Supplementary Tables 1 and 2). Specifically, we found 100 pathways that were enriched in both tumor types. Although GSEA alone is a powerful tool to identify pathways of genes that potentially play a role in different tumor types, we were also interested in alternative methods. For the identification of differentially expressed gene sets without resorting to any precompiled pathway or gene set information we assembled a network of experimentally obtained pairwise molecular phosphorylation events, protein–protein and protein–DNA interactions. Applying PinnacleZ [26], a Cytoscape plugin [27], to a network of 103,966 interactions between 11,969 human proteins we found 89 differentially expressed subnetworks in the HF data and 32 in the GMDI data ($P < 0.05$, Supplementary Tables 3 and 4) (Fig. 1b). To test the validity of our hypothesis that both enriched pathways and differentially expressed subnetworks can reliably distinguish between different tumor types, we applied a Support-Vector Machine approach [29], utilizing gene sets as classifiers. Specifically, we transformed our gene expression profiles of GBMs and oligodendrogliomas into a matrix of Z-scores [26]. For each set of genes M_k we calculated an activity score in sample j,

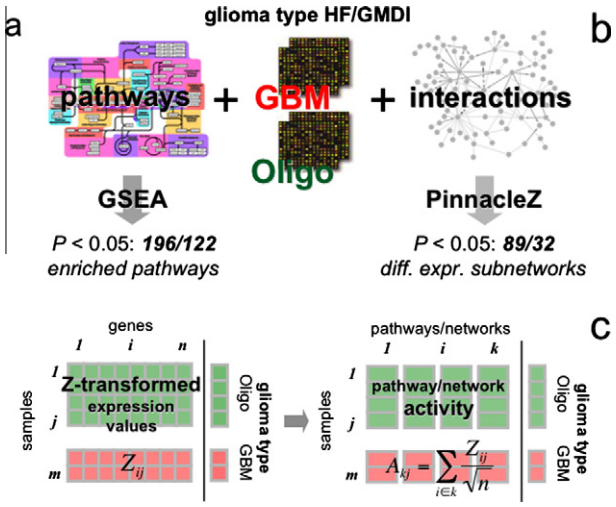

Fig. 1. We utilized gene expression samples of Glioblastoma Multiforms (GBM) and Oligodendrogliomas (Oligo) from two totally disparate data sets (HF and GMDI). (a) Using Gene Set Enrichment Analysis (GSEA), we found 196 and 122 pathways that are enriched in one glioma type in HF and GMDI data, respectively. (b) Applying the PinnacleZ algorithm we found 89 differentially expressed subnetworks in the HF samples and 32 networks in the GMDI data. (c) In order to score each subnetwork/pathway, we determined a sample-specific *Z*-score for each gene by averaging over all samples and defined the activity of each pathway/network as an average of all involved genes.

A_{kj} (Fig. 1c). Utilizing enriched pathways, each sample was represented by an *x*-dimensional vector, holding the activity scores of all *x* pathways in the underlying sample. Running a 3-fold cross-validation 1000 times, we observed that using 196 enriched pathways as classifiers allowed us to obtain a high percentage of correct predictions in the HF data set (Fig. 2a). Determining the prediction rates for oligodendrogliomas and GBMs separately, however, we observed that the high overall rate of correct predictions was mainly secondary to the correct prediction of the GBM group. By contrast, the ability of the classifiers to correctly predict oligodendrogliomas dropped significantly. Since the activity of each enriched pathway was defined as the average of each genes expression, we reason that the presence of genes that do not

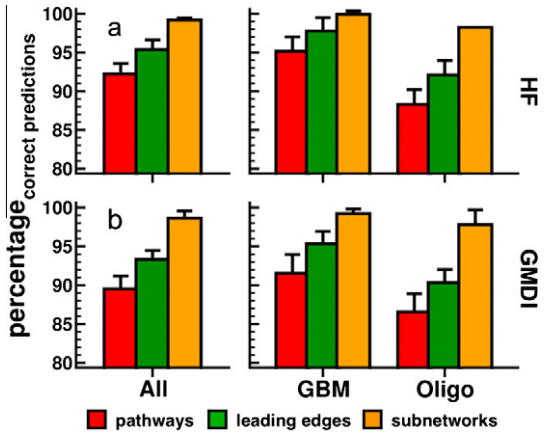

Fig. 2. Repeating 3-fold cross-validation 1000 times with a support-vector machine approach, we found in (a) that enriched pathways in HF data allow a high overall rate of correct predictions. However, the performance dropped if we predicted Oligodendrocytes separately. Accounting for genes that drive the pathways enrichment (i.e. leading edges) we observed a significant performance improvement. Compared to pathway specific results, differentially expressed subnetworks significantly outscored enriched pathways as classifiers. (b) Utilizing GMDI data, we found similar results, indicating the superior performance of differentially subnetworks as classifiers.

contribute significantly to the enrichment of its pathway might be responsible for the observed diminished performance. To test this hypothesis, we determined the activity of each pathway based only on the genes that significantly enhanced the enrichment of its pathways (i.e. 'leading edges'). As demonstrated in Fig. 2a, we found that the use of leading edges significantly improved the performance of our pathway-based classifiers. Analogously, we checked the performance of classifiers composed of 89 differentially expressed subnetworks. By contrast, we observe that such differentially expressed subnetworks significantly outscored enriched pathways, allowing for correct prediction rates up to 99% in the HF data set (Fig. 2a). While the performance of enriched pathways dropped considerably in predicting the oligodendroglioma group, subnetworks appeared to predict both tumor types equally well, a result we found with the GMDI data set as well (Fig. 2b).

3.2. Genomic alterations

The presence of many enriched pathways in GBMs and Oligodendrogliomas and numerous genomic alterations in most gliomas suggested the possibility that many of these pathways are prime targets of genomic alterations. In order to test this hypothesis, we collected a list of 1979 genes associated with chromosomal abnormalities in at least 15% of gliomas. Applying a Fisher's exact test (nominal $P < 0.05$), we found that 19.9% and 16.4% of all enriched pathways in the HF and GMDI data set, respectively, are significantly populated with genes that have at least one indication of a genomic alteration. These numbers increase to 23.0% and 25.5% when utilizing the leading-edge genes for the HF and GMDI data sets, respectively. Genes that appear in many different enriched pathways might play an important role in aberrant cell signaling and may be enriched for targets of genomic alterations that support the emergence of the malignant phenotype. To test this hypothesis, we connected enriched pathways that share such genes. Applying a threshold of a nominal $P < 0.05$ obtained from an adapted two-tailed Fisher's exact test, we found such genes in an assembly of degradation pathways that are involved in cell-cycle regulation and in prominent signaling pathways (Fig. 3). While genomically altered proteasome subunits drive the enrichment signal in degradation pathways we also found that the tightly connected toll-like receptor, BCR signaling and IL1r signal transduction pathways were strongly affected by genomically altered genes. Taken together, the mentioned pathways pool an array of important signaling proteins such as MAP-kinases, TNF receptor associated factor TRAF6, the transcription factor NF-kB and its inhibitor I-kB kinase as well as FOS and JUN, genes that form the AP-1 early response transcription factor. In the GMDI data set, we found 5 out of 32 (15.6%) strongly overexpressed subnetworks in GBMs that are significantly enriched (two-tailed Fisher's test, $P < 0.05$) with genomically altered genes. Analogously, 8 out of 89 subnetworks (9.0%) were found enriched with genomically altered genes in the HF data (Supplementary Fig. 1). Such webs largely revolve around the epidermal growth factor receptor (EGFR) and BCL3, both of which can have genomic alterations in gliomas (Fig. 4). The latter proto-oncogene functions as a transcriptional coactivator through its association with NF-kappa B homodimers, an activity that is reflected by the strong presence of protein interactions between these genes in Fig. 4 and enriched pathways in Fig. 3.

Another important step that would support the applicability of enriched pathways and differentially subnetworks as classifiers would be the ability to predict correct classes if the procedure were trained on the HF data set, tested on the GMDI data set and *vice versa*. In Fig. 5a, we generally found that the leading edges and subnetworks based classifiers provided the best classification accuracy, while training on GMDI and testing on HF data largely

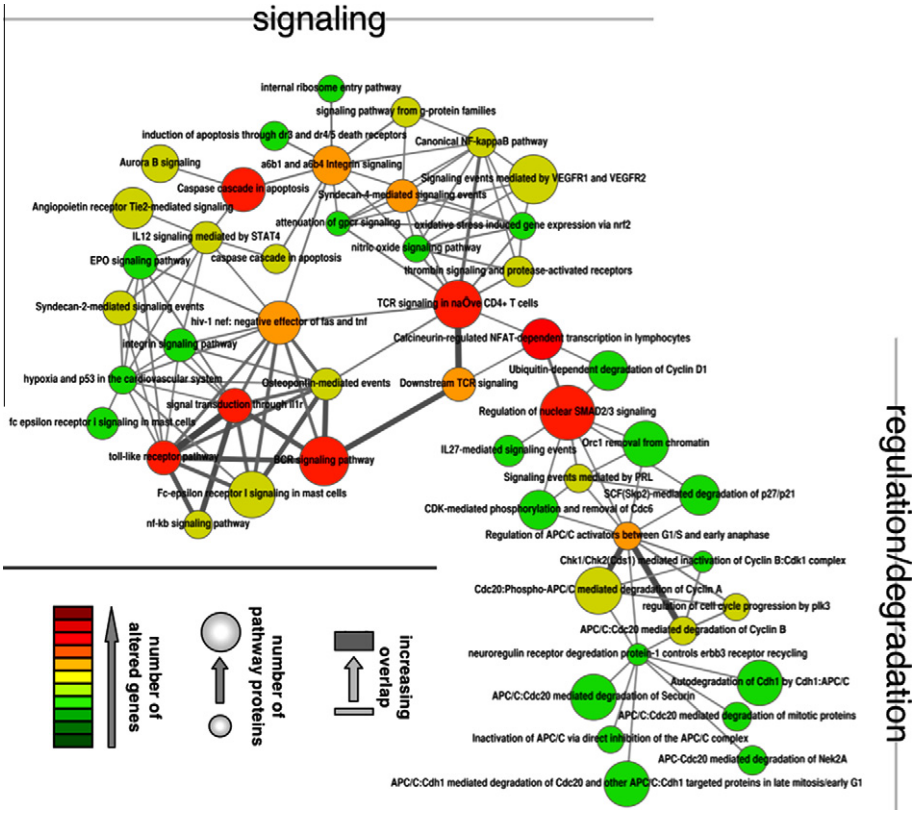

Fig. 3. Focusing on enriched gene sets, we connected pathways if they significantly share genes with evidence of a genomic alteration (nominal $P < 0.05$). We mostly found degradation pathways involved in cell-cycle regulation activities. In addition, we observed numerous prominent signaling pathways that were strongly composed of genes with genomic alterations in gliomas.

outperformed training on the HF data and testing on the GMDI data set. Focusing on oligodendroglioma specific results, we observed that all classification procedures dropped in accuracy quite dramatically, while the GBM specific performance remained strong (Fig. 5b).

3.3. Selection of optimized pathways and subnetworks

Up to this point we had utilized all enriched pathways and differentially expressed subnetworks to determine the classifiers. Since the identified subnetworks had been determined independently from each other, however, we wondered if certain feature combinations would improve the general performance of the classifiers, and specifically the prediction accuracy of the Oligodendroglioma samples. The iterative relief algorithm (I-Relief) offers a computational framework to identify such combinations by weighting the underlying features in an iterative manner [31,32]. Running through a ranking of all enriched pathways or subnetworks according to their weight, we iteratively added a new feature. Checking the prediction accuracy of each subset of features, we defined the group of features with the highest prediction accuracy as the optimal subset of pathways or subnetworks that allow the best discrimination between GBMs and Oligodendrogliomas (Supplementary Fig. 2). With HF data as the training and GMDI data as the testing set, we found that the prediction accuracy of the 11 top-ranked pathways clearly outscored all 196 enriched pathways (Fig. 5a). Analogously, 21 leading edge sets performed significantly better than all their counterparts, and 7 subnetworks reached equivalent levels of accuracy to all 89 differentially expressed sub-webs. In the opposite case, training on GMDI and testing on the HF data set also led to an increase in the overall prediction accuracy with relatively low numbers of selected pathways and subnetworks (Fig. 5a). More importantly, however, the Oligodendroglioma specific prediction accuracy strongly increased in both cases as well, favouring GMDI over HF data as training rather than a testing set (Fig. 5b).

Utilizing 3-fold cross-validation in each data set separately, we also tested the prediction accuracy of the corresponding optimal pathways and subnetworks sets. In Supplementary Fig. 3a, we found that the general and type specific prediction accuracy was generally high with optimized subnetworks and pathways using the HF data set. Using the GMDI data set, optimized pathways classified GBMs well while we observed a slight drop in the accuracy of predicting the group of oligodendrogliomas. While subnetworks seemed to work equally well with both glioma types, our result suggested that the low performance in the prediction of oligodendrogliomas is largely a question of the optimized subsets of pathways and subnetworks, yet are slightly influenced by the intrinsic characteristics of the underlying expression data sets.

The identification of optimized subsets of pathways and networks suggested potential biological significance. Pooling of the 6 optimal subnetworks obtained with the GMDI data set resulted in a network that revolves around prominent tumor related genes such as NFKB1, VEGF and BCL3 (Fig. 5b), genes that appeared in the unified web of the 7 most discriminative networks in the HF data as well (Supplementary Fig. 4). In Supplementary Fig. 5, we also show connections between the corresponding 21 optimal leading edge sets in HF and GMDI data, respectively. Roughly, both networks revolved around prominent signaling and apoptotic pathways, indicating that SHC1, ITGB1 and CASP3 provide crosstalk between these enriched and highly discriminative pathways.

As a corollary to our initial hypothesis that enriched pathways and subnetworks are prime-targets of genomic alterations, we hypothesized that the tendency of such gene sets to be primarily

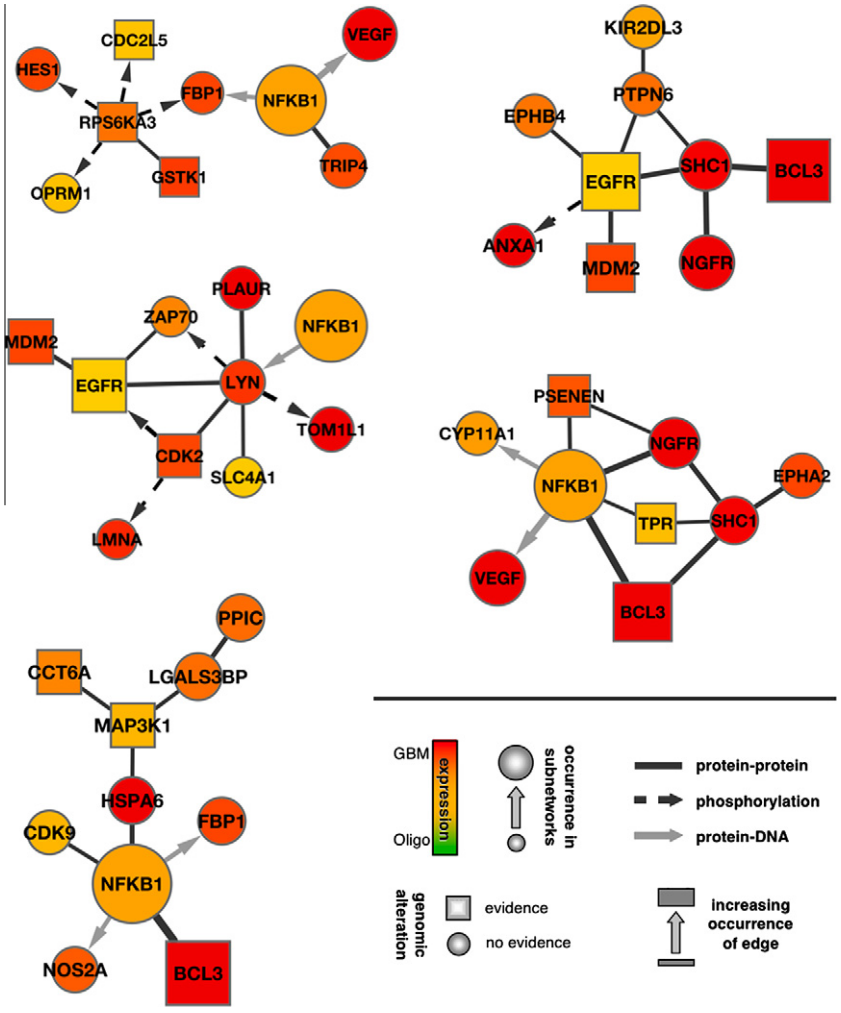

Fig. 4. Considering 89 differentially expressed subnetworks in the HF data set, we found 5 subnetworks that were significantly enriched with genes that showed a genomic alteration (Fisher exact test, $P < 0.05$). Specifically, such subnetworks strongly revolved around NFKB1 and EGFR.

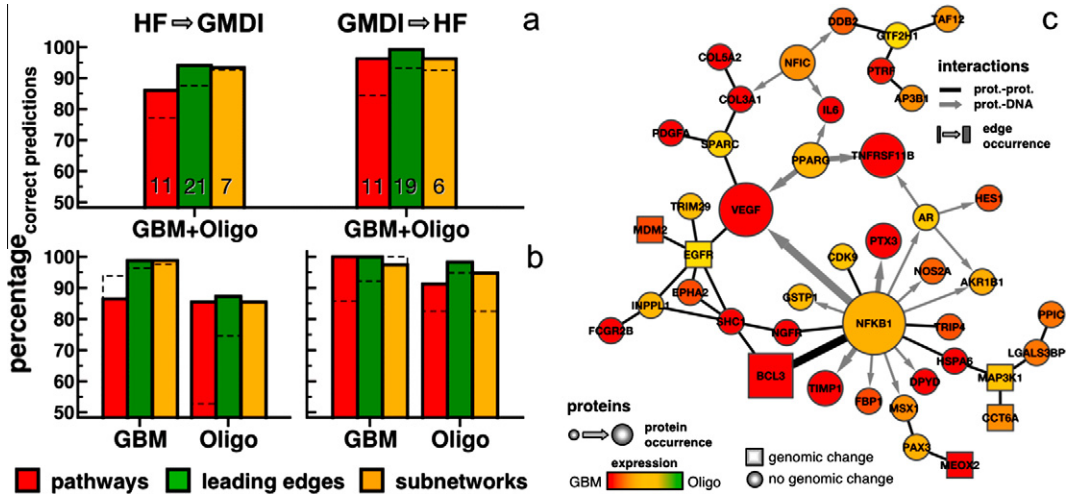

Fig. 5. (a) Training a SVM-based classification procedure on enriched pathways in HF data and testing the classifier on GMDI data, we found the best performance with 11 pathways, that clearly outscore the classification results using all enriched pathways (dashed line). Specifically, we ranked each pathway according to its contribution to the discrimination between GBM and Oligodendrogliomas and chose those that obtained the highest prediction accuracy. Analogously, we found similar results with 21 sets of leading edges and 7 subnetworks. In the opposite case, training in GMDI data and testing on the HF set increased the overall accuracy of the prediction with comparable numbers of pathways/subnetworks. In (b), we found that the glioma specific prediction accuracy depends on the quality of the underlying training set. While Oligo-specific predictions based on the HF set showed an advantage in predicting GBMs, training on the GMDI sets clearly eradicated this bias. In (c) we pooled the 6 subnetworks obtained with the GMDI data set and found that such a network revolved around prominent tumor related genes. In particular, this network was strongly governed by NFKB1, VEGF and BCL3, although only the latter two genes were strongly overexpressed in GBMs.

enriched with altered genes will increase. Out of 11 optimized pathways, we found one pathway (9.1%) to be significantly enriched with genomically altered genes ($P < 0.05$) utilizing HF data. Similarly, we found 4 out of 21 leading edge sets that have significant enrichment levels (19.0%), while we found no such enriched subnetworks. Analogously, we found 2 out of 11 pathways (18.2%), one out of 19 (5.3%) and one out of 6 subnetworks (16.7%) to be significantly enriched with genomically altered genes. Compared to the enrichment rates of all pathways and subnetworks, genomic alterations alone cannot convincingly explain the biological basis for the discriminative power of optimized sets of pathways and subnetworks.

4. Conclusions

Current clinical glioma classification schemes are based on histopathological observations and suffer from intra-observer subjectivity, poor prognostic or therapeutic predictive potential and offer little biological insights. Newer classification schemas based on gene expression are more objective and offer hints of biology, but many of the individual genes identified as classifiers exist as isolated markers with questionable relevance to tumor biology. The determination of enriched pathways and differentially expressed subnetworks in GBMs compared to oligodendrogliomas allowed us to find gene sets that can be used to reliably distinguish between underlying tumor types. Such a strategy may in fact have more biological meaning than using single probesets/genes in isolation since genes mostly mediate their biological functions as large assemblies rather than as single entities. Therefore, the genetic/protein associations found within differentially expressed pathways and subnetworks may help identify genes/proteins with important biological relevance to the pathogenesis of those tumors. Merely focusing on single genes alone would have otherwise missed such pathways and subnetworks.

Although we found that precompiled gene sets, such as pathways, have value in distinguishing classes of expression profiles, they are outperformed by a more local search. This strategy involved discovering differentially expressed subnetworks, where we scanned the vicinity of an individual gene in an interaction network for the purpose of finding interaction partners that add to the discriminative power of a single subnetwork. While each pathway carries genes that do not significantly contribute to its enrichment in one tumor class, this disadvantage can be partially compensated for by focusing on leading edges, genes that indeed are responsible for the observed overrepresentation of the underlying pathway. Such simple steps allowed us to identify an array of discriminative features providing high prediction accuracy by mutually training and testing on data sets of different origin. While we initially based the classification procedure on all individually enriched pathways and differentially expressed subnetworks, we refined our procedure by choosing optimal subsets of such features, allowing better classification performance as an ensemble. Specifically, we found that a relatively low number of pathways and subnetworks does not only drastically improve the classifiers performance but also allows us to compensate for the seemingly lower prediction accuracy of oligodendroglioma samples.

The major power of this approach over the single gene classifier approach is that identified enriched pathways may carry valuable biological information. We observed that a number of the highly discriminatory signaling and regulation pathways were significantly enriched with genomically altered genes in GBMs, suggesting their potential role in the biology of the underlying tumor type. In addition, we found subnetworks that were significantly populated with genomically altered genes largely revolving around hubs that play important roles in signal transduction and transcriptional regulation. The central location of such genes within these networks, compounded by their occurrence within frequently altered areas of the glioma genome, may hint to their potential importance in glioma pathogenesis. Since pathways carry many genes that are not significantly enriched in one tumor type, the determination of subnetworks works in a more selective way by accounting for only those genes that strongly help to discriminate the two classes. Indeed, we found major subnetworks revolving around prominent glioma related genes such as NFKB1, VEGF and BCL3.

Based on these observations, one might therefore assume that the strategy of finding subnetworks would more likely identify genes that are genomically altered in gliomas. Curiously, however, we did not find any evidence that differentially expressed subnetworks accumulate more genomically altered genes than enriched pathways. In fact, the enrichment of genomically altered genes in optimized subsets did not differ compared to all pathways and subnetworks, suggesting that the presence of altered genes is largely not a criterion for the discriminative power of optimized pathways and subnetworks. In turn, one might speculate that the paucity of genomically altered genes in our subnetworks points to the possibility that genes in the subnetworks, albeit discriminative, are the result of genomically altered regulators that exert their influence in a way that is not detectable by focusing on local interaction networks alone. Another explanation might be that we possibly underestimated the number of genomically altered genes in our analysis. Specifically, we determined altered genes using frequent chromosomal number alterations in gliomas. However, single nucleotide polymorphism (SNP) generated genomic arrays only identify rather large genomic alterations. Therefore, some of the enriched genes that were not designated as "genomically altered" might have undergone smaller (i.e. single base pair) mutations in their coding or upstream promoter sequences. Such small changes result in altered mRNA stability and/or altered transcriptional activity potentially influencing our results.

Since the level of expression of genes can be explained by genomic alterations in many (although not all) cases, a refinement of our methodology would be the identification of biologically important genes that are regulated by genomically altered genes. Although our approach of identifying enriched gene sets and differentially expressed subnetworks moves us in that direction, such an approach is still dependent on well-annotated precompiled gene sets and/or local gene–gene and protein–protein interactions that are largely left to be uncovered. Furthermore, difficulties in distinguishing passenger from driver genetic mutations and alterations would complicate such an approach. The true value of this approach, ultimately awaits vigorous cross-validation using desperately needed outside data sets and experimental validation of the importance of the identified pathways, subnetworks and the specific genes and proteins.

Our approach offers a proof of principle that pathway and subnetworks reliably allow classification and potential insights into the underlying biology of those tumors. In addition, such subnetworks and pathways might help to determine distinct subsets of glioblastomas and oligodendrogliomas and/or help to distinguish clinical outcomes.

Such insights are important for both understanding tumor biology and for the selection of appropriate molecularly targeted therapy for individual tumors as we head into an age of personalized medicine.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2010.08.011.

References

[1] Harras A. Cancer rates and risks. 4th ed. Washington, DC: US Department of Health and Human Services, Public Health Service, National Institutes of Health; 1996.

[2] Kitange GJ, Templeton KL, Jenkins RB. Recent advances in the molecular genetics of primary gliomas. Curr Opin Oncol 2003;15:197–203.

[3] Rich JN, Sathornsumetee S, Keir ST, Kieran MW, Laforme A, Kaipainen A, et al. ZD6474, a novel tyrosine kinase inhibitor of vascular endothelial growth factor receptor and epidermal growth factor receptor, inhibits tumor growth of multiple nervous system tumors. Clin Cancer Res 2005;11:8145–57.

[4] Liang Y, Diehn M, Watson N, Bollen AW, Aldape KD, Nicholas MK, et al. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. Proc Natl Acad Sci USA 2005;102:5814–9.

[5] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.

[6] Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 2006;9:157–73.

[7] Li A, Walling J, Ahn S, Kotliarov Y, Su Q, Quezado M, et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. Cancer Res 2009;69:2091–9.

[8] Miller CL, Diglisic S, Leister F, Webster M, Yolken RH. Evaluating RNA status for RT-PCR in extracts of postmortem human brain tissue. Biotechniques 2004;36:628–33.

[9] Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999;21:20–4.

[10] Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA 2001;98:31–6.

[11] Kotliarov Y, Steed ME, Christopher N, Walling J, Su Q, Center A, et al. High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. Cancer Res 2006;66:9428–36.

[12] Kotliarov Y, Kotliarova S, Charong N, Li A, Walling J, Aquilanti E, et al. Correlation analysis between SNP and expression arrays in gliomas identify potentially relevant targets genes. In: Cancer Res; 2008.

[13] Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153–8.

[14] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature 2005;437:1173–8.

[15] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein–protein interaction network: a resource for annotating the proteome. Cell 2005;122:957–68.

[16] Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, et al. Large-scale mapping of human protein–protein interactions by mass spectrometry. Mol Syst Biol 2007;3:89–99.

[17] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, et al. IntAct – open source resource for molecular interaction data. Nucleic Acids Res 2007;35:D561–5.

[18] Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 2008.

[19] Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al. MINT: the Molecular INTeraction database. Nucleic Acids Res 2007;35:D572–4.

[20] Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, et al. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res 2004;32:D497–501.

[21] Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res 2007;35:D137–40.

[22] Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, Miron IM, et al. Systematic discovery of in vivo phosphorylation networks. Cell 2007;129:1415–26.

[23] Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, Bork P, et al. NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic Acids Res 2008;36:D695–9.

[24] Diella F, Gould CM, Chica C, Via A, Gibson TJ. Phospho.ELM: a database of phosphorylation sites – update 2008. Nucleic Acids Res 2008;36:D240–4.

[25] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. Nucleic Acids Res 2009;37:D674–9.

[26] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol 2007;3:140.

[27] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.

[28] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.

[29] Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci USA 2007;104:4337–41.

[30] Francesconi M, Remondini D, Neretti N, Sedivy JM, Cooper LN, Verondini E, et al. Reconstructing networks of pathways via significance analysis of their intersections. BMC Bioinformatics 2008;9(Suppl. 4):S9.

[31] Sun Y. Iterative RELIEF for feature weighting: algorithms, theories, and applications. IEEE Trans Pattern Anal Mach Intell 2007;29:1035–51.

[32] Sun Y, Goodison S, Li J, Liu L, Farmerie W. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics 2007;23:30–7.