 doi:10.20944/preprints201908.0100.v1



RESEARCH

# KARL: Knowledge Augmented Rule Learning for Informed Biomarker Discovery

Henry A. Ogoe[1,2†], Mahbaneh Eshaghzadeh Torbati[3†] and Vanathi Gopalakrishnan[1,3,4,5*]

*Correspondence: vanathi@pitt.edu
[1]Department of Biomedical Informatics, University of Pittsburgh, 5607 Baum Bld, Pittsburgh, PA 15206, USA
Full list of author information is available at the end of the article
†Equal contributor

**Abstract**

**Background:** Ongoing molecular profiling studies enabled by advances in biomedical technologies are producing vast amounts of 'omic' data for early detection, monitoring and prognosis of diverse diseases. A major common limitation is the scarcity of biological samples, necessitating integrative modeling frameworks that can make optimal use of available data for disease classification task. Related data sets are often available from different studies, but may have been generated using different technology platforms. Thus, there is a critical need for flexible modeling methods that can handle data from diverse sources to facilitate discovery of robust biomarkers that underlie disease regulatory processes.

**Results:** In this paper, we introduce a novel framework called Knowledge Augmented Rule Learning (KARL), which incorporates two sources of knowledge, domain and data, for pattern discovery from small and high-dimensional datasets, such as transcriptomic data. We propose KARL as a transfer rule learning framework in which knowledge of domain is transferred to the learning process on data in order to 1) improve the reliability of the discovered patterns, and 2) study the knowledge of the domain when used along with data for modeling. In this work, we generated KARL models on gene expression datasets for five types of cancer, including brain, breast, colon, lung, and prostate. As our knowledge of the domain, we used the Ingenuity Knowledge Base (IKB) to extract genes related to hallmarks of cancer and annotated these prior relationships before learning classifiers from these datasets.

**Conclusions:** Our results show that KARL produces, on average, rule models that are more robust classifiers than the baseline without such background knowledge, for our tasks of cancer prediction using 25 publicly available gene expression datasets. Moreover, KARL helped us learn insights about previously known relationships in these gene expression datasets, along with new relationships not input as known, to enable informed biomarker discovery for cancer prediction tasks. KARL can be applied to modeling similar data from any other domain and classification task. Future work would involve extensions to KARL to handle hierarchical knowledge to derive more general hypotheses to drive biomedicine.

**Keywords:** cancer biomarker discovery; gene expression data; Ingenuity Knowledge Base (IKB); transfer learning; interpretable classification rules

## Background

The advent of high-throughput genomic techniques like the microarray [1] and next generation sequencing [2–4] have enabled the profiling of vast amounts of transcriptomic data with increasing accuracy over the past decades [5, 6]. This type of data is a collection of all RNA transcripts, including both protein non-coding RNAs and

 doi:10.20944/preprints201908.0100.v1



coding mRNAs. The mRNA transcripts at particular time points can give an indication of the functional role of a gene, the underlying mechanism of a disease, or a potential drug response [1,7,8]. Having such knowledge, a myriad of computational techniques have been proposed by biomedical data scientists to analyze such 'omic' profiles, ranging from simple statistical testing to more sophisticated bioinformatics and machine learning methods.

Predictive disease modeling using transcriptomic data typically yields models that can be roughly categorized into two main groups. The first group includes models obtained from popular methods such as logistic regression, artificial neural networks (ANN), and support vector machines (SVMs), wherein a mathematical function is learned as the predictor. The second group includes models represented as interpretable decision trees or classification rules, obtained via inductive learning of symbolic descriptions. The former group includes models with relatively high predictive performance, but the majority of them (e.g., ANN, SVMs) are hard to interpret. On the other hand, the latter category consists of highly interpretable models with lower but acceptable predictive performance. The high interpretability of models of the second category makes them suitable for the task of pattern discovery.

Models in both of these categories can be adversely affected by two common characteristics of the transcriptomic data: (1) the small sample size (tens to hundreds), and (2) the large number of variables (ranging from hundreds to several thousand.) The models learned on such sparse high-dimensional data are likely to not generalize well due to the lack of a strong statistical support for their predictions, and also due to sampling bias and variability. Even if similar Case-Control studies are conducted across different laboratories, the heterogeneity of sample sources and experimental protocols of different datasets may prevent learning of robust models on the union of the resultant datasets.

The Transfer Rule Learner (TRL) framework has been developed to take on this challenge of modeling sparse biomarker data for pattern discovery by using transfer learning [9]. Transfer learning is the broad concept of using any type of knowledge/information from any possible type of source for enhancing the learning process on the target task [10]. TRL uses as its main classification engine, Rule Learner (RL) [11], which adopts a general-to-specific search strategy to learn its prediction model as a set of classification rules. Each classification rule is represented as IF-condition THEN-consequent, where the Condition is expressed as a logical AND of a small subset of variable-value pairs (predictors), and the Consequent refers to the value of the Class variable. As the transfer learning strategy, TRL automatically accepts the classification rules, prior rules, learned from a source dataset, when learning from a target dataset. TRL has been applied to the analyses of two related proteomic mass spectral datasets obtained from separate institutions. The results showed that positive transfer of knowledge is observable through increased predictive performance, when the target data is sparse relative to the source dataset.

TRL was the first application of transfer learning of classification rules to biomarker discovery, but it was limited to the homogeneous source and target datasets. Its next version, the TRL-FM [12], utilized the notion of functional modules to expand TRL to be applicable to heterogeneous source and target datasets. In TRL-FM, the functional modules provide a general mapping between the source

 doi:10.20944/preprints201908.0100.v1



and target datasets. TRL-FM, which was used for gene expression datasets, used the Gene Ontology as the domain background knowledge to cluster together genes within the same biological process, thereby mapping and expanding the knowledge that was transferred between the source and target to include additional prior rules with functionally similar genes. The results showed that TRL-FM outperformed six state-of-the-art traditional machine learning systems.

In both TRL and TRL-FM approaches, a related dataset is selected as the source of transfer, which makes the quality of transfer be dependent on the effectiveness of the source data for the target task. This may result in *negative* transfer in which transfer deteriorates the performance of the baseline model. The negative transfer can be the result of noise in data or difference in variable distribution between datasets. KARL is a general framework in which a more reliable source of knowledge is designed to be used as the source of transfer, that is, subjective domain knowledge. In the KARL framework, both the subjective and objective knowledge is used in the learning process of the RL model. While the objective knowledge comes from the target dataset, the subjective knowledge would be a reliable knowledge of domain sourced from experts, literature, or domain knowledge bases; or it can be an experts' hypothesis needed to be studied or verified over data. We propose KARL as:

1 A framework in which subjective domain knowledge can be used in the learning process as a direct source of transfer, to assist the learning of models from sparse data and perform informed discovery on a target dataset.
2 A framework by which the knowledge of domain and data can be tracked, visualized, estimated, and studied in the transfer process and final model, in order to provide both insights over the interactions between sources, and an easy way to verify whether source knowledge patterns are found in the target data via annotations.

KARL accepts the knowledge of domain abstracted as a table of variables, lookup table. Then, it generates the set of all possible single-variable rules from this table, called prior rules. The prior rules are then used in two transfer learning strategies, resulting in two sets of rules: Rule Model with Prior Rules and Evidence-based Prior rules. These two rule sets are then combined and constitute the KARL rule model. In the first strategy, RL with Priors ($RL_P$) as an expanded version of RL accepts the prior rules as its initial search point and generates Rule Model with Prior Rules as the rule model learned on the target dataset. This strategy may help $RL_P$ to start from a better and more feasible point in the search space, which is provided from the knowledge of the domain. In the second strategy, the set of the prior rules, pruned on the datasets using the criteria defined by the user, are generated as Evidence-based Prior rules which is directly transferred to the final KARL model. This transfer learning strategy would improve the design by using the knowledge of domain which may not be extracted from the small datasets. These two rule sets are then combined and constitute the KARL rule model.

In the KARL framework, the rules are marked by their source of origin, domain or data, and can be tracked during the learning process. This provides KARL with the ability to visualize the transfer, which to the best of our knowledge is a new functionality that is not provided in any other transfer rule learning framework. Tracking the rules also provides the KARL model with information on the interaction of sources and the estimation of their involvement in the learning process.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1



These properties enable the scientists to study their subjective knowledge of interest along with data using KARL. This may also give scientists more insights into their hypothesis over their markers of interest, as the class discriminative markers might be applicable to specific subgroups of data in the final model.

In this study, we have used our KARL framework to predict five different types of cancer, including brain, breast, colon, lung, and prostate from the gene-expression datasets. As our choice of background domain knowledge for cancer, we have selected the Ingenuity® Knowledge Base (IKB). The IKB contains evidence-based domain knowledge in the form of gene-interaction networks. We have used the IKB to extract those variables of domain that are associated with the general hallmarks of cancer, as our domain knowledge of interest. Three dominant properties of cancer are used as its hallmarks and include: the faulty control of the cell cycle, the faulty control of cell death, and the invasiveness and metastatic capabilities. We propose that our knowledge of domain as the table of those gene-expression variables showing the general hallmarks of cancer, will improve the predictive rule learning on cancer prediction. In this paper, we demonstrate that the knowledge discovery process in the rule learning for cancer prediction can be augmented with such knowledge of domain, and test the hypothesis that *knowledge augmented rule learning with KARL, using the general hallmarks of cancer produces, on average, rule models that are more robust than baseline RL without any background knowledge, using 25 publicly available gene expression datasets on cancer prediction, ten-fold cross-validation performance measures, and the assessment of model Consistency and Completeness.* We also hypothesize that *the KARL framework will give us more insight on our knowledge of domain or the general hallmarks of cancer, by visualizing the interaction of sources in its learning process.*

## Results

In this section, we first describe our experimental design along with the data and evaluation metrics developed and used. Then, in the next subsections, we report the details and the results from our experiments.

### Experimental Design

We designed our experiments to test the two properties that we have proposed for KARL: 1) involving subjective domain knowledge in the learning process in order to improve the predictive performance or generality of the rule learning models, 2) visualizing, estimating, and studying the interaction of the domain and data in the final model. For our experiments, we used publicly available gene expression data sets from the Gene Expression Omnibus (GEO) repository [13]. This data, described in Table 1, contains 25 datasets, consisting of 5 each from 5 different cancer types (i.e., brain, breast, colon, lung, and prostate).

We have designed KARL to accept the knowledge of domain as an input table of variables, lookup table. For our cancer prediction study, we proposed that the lookup table contain those variables of the domain that are associated with general hallmarks of cancer. KARL uses the transfer learning approaches to involve the prior rules built upon these variables in its learning process. In the transfer learning parlance, the concept of *positive* and *negative* transfer is used for evaluating the practicality of the transfer. Positive\negative transfer is said to have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1



Table 1: Description of gene expression datasets. Series ID = GEO accession number, ID = The dataset pointer referred to in Figure 1.

| Disease | Series ID, Source | ID | Platform | No. of Samples (Tumor/Normal) | No. of Variables |
|---|---|---|---|---|---|
| Brain Cancer | GEO16011, [14] | A | HG-U133 Plus2, [14] | 175 (159/16) | 17332 |
| | GEO1993, [15] | B | HG-U133A, [15] | 58 (39/19) | 12501 |
| | GEO4271, [16] | C | HG-U133A,B | 100 (76/24) | 28168 |
| | GEO4290, [17] | D | HG-U133 Plus2 | 100 (81/19) | 20185 |
| | GEO4412, [18] | E | HG-U133A,B | 85 (59/26) | 28168 |
| Breast Cancer | GEO10780, [19] | F | HG-U133 Plus2 | 185 (42/143) | 20156 |
| | GEO15852, [20] | G | HG-U133A | 86 (43/43) | 12501 |
| | GEO29431, [21] | H | HG-U133 Plus2 | 66 (54/12) | 20156 |
| | GEO42568, [22] | I | HG-U133 Plus2 | 121 (104/17) | 20156 |
| | GEO7904, [23] | J | HG-U133 Plus2 | 62 (43/19) | 20156 |
| Colon Cancer | GEO10715, [24] | K | HG-U133 Plus2 | 30 (19/11) | 20156 |
| | GEO20916, [25] | L | HG-U133 Plus2 | 70 (36/34) | 20185 |
| | GEO23878, [26] | M | HG-U133 Plus2 | 59 (35/24) | 20185 |
| | GEO24514, [27] | N | HG-U133A | 49 (34/15) | 12501 |
| | GEO9348, [28] | O | HG-U133 Plus2 | 82 (70/12) | 20185 |
| Lung Cancer | GEO10072, [29] | P | HG-U133A | 107 (58/49) | 12501 |
| | GEO18842, [30] | Q | HG-U133 Plus2 | 91 (46/45) | 20156 |
| | GEO19188, [31] | R | HG-U133 Plus2 | 156 (91/65) | 20156 |
| | GEO19804, [32] | S | HG-U133 Plus2 | 120 (60/60) | 20156 |
| | GEO7670, [28] | T | HG-U133A | 66 (39/27) | 12501 |
| Prostate Cancer | GEO17951, [33] | U | HG-U133 Plus2 | 137 (68/69) | 20185 |
| | GEO32448, [34] | V | HG-U133 Plus2 | 80 (40/40) | 20156 |
| | GEO46602, [35] | W | HG-U133 Plus2 | 50 (36/14) | 20156 |
| | GEO6956, [36] | X | HG-U133A-2 | 89 (69/20) | 12501 |
| | GEO82188, [37] | Y | HG-U133A | 136 (65/71) | 12501 |

occurred on a dataset, if the model of a framework with transfer learning outperforms\underperforms the model learned on the same framework without the transfer [10]. Accordingly, for showing the practicality of our choice of domain knowledge and transfer learning strategies, as well as, testing our first property of KARL, we compare KARL to RL. In our experiments in 'RL Evaluation', we first report the predictive performance of RL and show that it is on average comparable to a range of state-of-the-art classifiers. Then, we evaluate KARL for Robustness (explained below) compared to RL, in the next two subsequent subsections. All our experiments estimate and report model performances over 10-fold cross-validation for each dataset.

In the last two subsections, 'Domain and Data Involvement in the KARL Model' and 'Analysis of Patterns', we elaborate upon our KARL rule models on cancer datasets to show the practicality of the second property of KARL for biomarker discovery purposes. We depict a snippet of our rule models and propose two metrics to estimate the contribution of sources, domain and data, in the KARL final model. We also ascertain whether intuitive and distinct rule patterns could be discovered across multiple KARL models for related studies (e.g., same cancer type).

*Evaluation Metrics*

Rl as an agnostic method, classifies samples as either known (one of the Class labels), or unknown. RL abstains from predicting those samples for which it is not confident in making a decision, so the labels for these samples are treated as unknown. This behavior of RL, requires us to evaluate our rule models not only based

 doi:10.20944/preprints201908.0100.v1



on their *Consistency* (prediction performance), but also based on their *Completeness* (coverage over samples). We define the notion of *Robustness* to indicate both Completeness and Consistency for rule models generated by any framework that uses RL as its learning algorithm, including KARL. We say that a model is *Robust*, if it is *Complete* and *Consistent*.

We assess the Completeness of our rule models with *Coverage* and *Abstentions*. Given the training data, Coverage is the fraction of the instances which logically satisfy the IF-condition part of at least one rule in the model. The higher the Coverage is (i.e., closer to one), the more Complete a rule model is on the train data. Given a set of test data, the model abstain from predicting an instance, if none of its rules cover it. The number of instances that the model abstains from predicting is referred to as Abstentions. The less a rule model abstains from making a prediction, the more Complete it is on the test data. We also measure the *Abstention rate* in our experiments to capture the percentage of test samples that were not predicted over the cross-validation folds. For assessing the Consistency of models, we use classification performance measures, including Sensitivity (SN), Specificity (SP), Balanced Accuracy (BAcc), and Accuracy (Acc), considering Case/tumor as the positive Class. In addition, we recalculated the Accuracy measure by considering Abstentions as misclassified samples and call it Accuracy including Abstention (AccAb).

We also define the Robustness measure for each rule in the RL model, as we collect the set of Robust rules as our final discovered patterns for the target task. For estimating the Robustness for a rule, we use the same definition like that of the rule model, as a rule is actually a single-rule model. Each rule of a model is accompanied with statistics which provide the information for calculating the rule's Robustness. This statistics provide CF and Coverage as the rule's measure of Consistency and Completeness, respectively (we elaborate more upon this topic in 'KARL Rule Model' subsection.) The experts of the domain can then utilize the provided Robustness as the rule's reliability measure, in order to select their patterns of interest for further studies.

### RL Evaluation

RL has been used in various prediction and biomarker discovery studies, [12,18,38–41], and has provided models with high predictive power. In this experiment, we want to show that RL also provides reliable models over our datasets of cancer and thus is an acceptable method as the foundation for KARL to build on. To this aim, we compare RL to other standard tree/rule learners and state-of-the-art classifiers implemented in the WEKA toolkit [42], on our 25 datasets, discretized and grouped according to the 5 cancer types. The standard tree/rule learners include: Decision Tree (C4.5), RIPPER, and PART, as well as the state-of-the-art classifiers include: Logistic Regression (LR), Support Vector Machines (SVMs), Naive Bayes (NB), and Random Forests (RF).

Tables 2 and 3 depict the predictive performance (BAcc, Accuracy) of these methods averaged over the five datasets for each cancer type. These results are mainly presented to show that RL models, on average, have acceptable BAcc (see Table 2) and Accuracy (see Table 3) when compared to the state-of-the-art classifiers and the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1



standard tree/rule learners. The supporting evidence can be found in the last rows of Tables 2 and 3. These rows show that RL performs on average on par or better than the three tree/rule learners on both measures. These rows also show that RL has acceptable average BAcc and Accuracy compared to the state-of-the-art classifiers, which have on average higher predictive performance, but lack *interpretability* to be used for pattern discovery in KARL. The same observations apply for each type of cancer, as seen in these comparisons. RL outperforms at least two of the other rule learners and has comparable performance to the state-of-the-art classifiers on all cancers except brain, using the BAcc measure (see Table 2). Brain cancer is a challenging dataset for RL which results in low Specificity and accordingly low BAcc measure. As observed in Table 3, RL outperforms at least one other rule learner and also shows acceptable results when compared to the state-of-the-art models on all cancers, using the Accuracy measure.

Table 2: Comparing RL with rule learning methods, including Decision Tree (C4.5), RIPPER, and PART and some state-of-the-art classifiers, including Logistic Regression (LR), Support Vector Machines (SVMs), Naive Bayes (NB), and Random Forests (RF) based on BAcc over 10-fold cross-validation.

| Disease | RL | C4.5 | RIPPER | PART | LR | SVMs | NB | RF |
|---|---|---|---|---|---|---|---|---|
| Brain | 63.2 | 72.6 | 73.5 | 73.2 | 69.0 | 74.3 | 79.9 | 73.3 |
| Breast | 83.4 | 81.9 | 84.0 | 81.1 | 86.6 | 88.1 | 85.1 | 83.4 |
| Colon | 88.1 | 83.2 | 87.6 | 83.2 | 88.1 | 94.2 | 93.8 | 94.7 |
| Lung | 94.9 | 94.4 | 92.4 | 94.5 | 93.6 | 97.5 | 96.4 | 96.8 |
| Prostate | 88.1 | 83.0 | 80.0 | 83.1 | 84.2 | 88.9 | 84.5 | 86.5 |
| Average | 83.6 | 83.0 | 83.5 | 83.0 | 84.3 | 88.6 | 87.9 | 86.9 |

Table 3: Comparing RL with rule learning methods, including Decision Tree (C4.5), RIPPER, and PART and some state-of-the-art classifiers, including Logistic Regression (LR), Support Vector Machines (SVMs), Naive Bayes (NB), and Random Forests (RF) based on Accuracy over 10-fold cross-validation.

| Disease | RL | C4.5 | RIPPER | PART | LR | SVMs | NB | RF |
|---|---|---|---|---|---|---|---|---|
| Brain | 83.3 | 82.5 | 82.8 | 84.1 | 84.9 | 88.8 | 85.3 | 88.6 |
| Breast | 87.7 | 85.0 | 86.6 | 85.5 | 87.2 | 91.8 | 89.1 | 88.5 |
| Colon | 88.2 | 87.7 | 88.5 | 87.7 | 86.6 | 95.0 | 94.5 | 95.7 |
| Lung | 92.9 | 94.4 | 92.2 | 94.5 | 93.3 | 97.4 | 96.3 | 96.8 |
| Prostate | 87.8 | 84.3 | 82.1 | 84.4 | 84.5 | 91.0 | 87.2 | 89.9 |
| Average | 88.0 | 86.8 | 86.4 | 87.2 | 87.3 | 92.8 | 90.5 | 91.9 |

## KARL Evaluation on Robustness

### *KARL Completeness*

Table 4 shows the average Coverage and Abstention rate statistics per disease using KARL. The Coverage shows the fraction of training samples covered by the model, and the Abstention rate denotes the percentage of the test samples for which the model abstained from making a prediction. While the Abstention rate ranges from 0 to 100, its values (AbsRate) are below 1% in Table 4, translating to almost zero Abstentions for brain and lung cancer test samples, and less than three Abstentions for colon cancer (explanation to follow). The colon cancer set results in the highest Coverage on average, with medians of 0.95 and 0.96 in normal and tumor examples, respectively. It also shows the least variation in Coverage that is [0.92, 0.97] and [0.92, 0.99] for normal and tumor samples, respectively. Brain cancer records the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1



worst average Coverage (median = 0.56 and 0.62 for Controls and Cases, respectively). Furthermore, the average Abstention rates ranges from 0.0 (brain and lung cancer) to 0.92 (colon cancer). The colon cancer datasets have the highest Abstention rate because there are fewer total number of samples overall to train on. Since cross-validation implies that every sample will be eventually appear as a test sample, the total number of test samples that are not predicted will be less than three. This number is obtained by adding up all the samples available for colon cancer as shown in Table 1, and rounding up the number that is closest to 1% of this total. The Completeness statistics of KARL on the entire 25 cancer datasets can be found in Additional File 2.

Table 4: Average Completeness rate on 10-Fold cross-validation per disease, using KARL approach. ConMin = minimum Coverage for Controls, ConMax = maximum Coverage for Controls, ConMdn = median Coverage for Controls, CaseMin = minimum Coverage for Cases, CaseMax = maximum Coverage for Cases, and CaseMdn = median Coverage for Cases, and AbsRate = %test samples not assigned a class.

| Disease | ConMin | ConMax | ConMdn | CaseMin | CaseMax | CaseMdn | AbsRate(%) |
|---|---|---|---|---|---|---|---|
| Brain | 0.39 | 0.73 | 0.56 | 0.46 | 0.79 | 0.62 | 0.00 |
| Breast | 0.65 | 0.84 | 0.76 | 0.64 | 0.82 | 0.72 | 0.22 |
| Colon | 0.92 | 0.97 | 0.95 | 0.92 | 0.99 | 0.96 | 0.91 |
| Lung | 0.76 | 0.98 | 0.85 | 0.83 | 0.97 | 0.91 | 0.00 |
| Prostate | 0.49 | 0.88 | 0.66 | 0.55 | 0.87 | 0.69 | 0.15 |

*KARL Consistency*

Table 5 represents the Consistency of KARL per disease type. Brain cancer recorded the worst average Accuracy including Abstention (i.e., AccAb = 83.59%), while lung cancer (i.e., AccAb = 94.79%) had the highest, followed closely by colon cancer (i.e., AccAb = 94.27%). The low BAcc from the brain cancer datasets, especially the low Specificity, could be attributed to a high degree of inherent heterogeneity in the instances. The Consistency statistics of KARL on the entire datasets can be found in Additional File 3.

Table 5: Average classification performance percentage on 10-fold cross-validation per disease, using KARL method. AccAb = Accuracy including Abstention, Acc = Accuracy, SN = Sensitivity, SP = Specificity, BAcc = Balanced Accuracy.

| Disease | AccAb | Acc | SN | SP | BAcc |
|---|---|---|---|---|---|
| Brain | 83.59 | 83.59 | 95.48 | 37.24 | 66.36 |
| Breast | 89.25 | 89.44 | 83.19 | 83.03 | 83.11 |
| Colon | 94.27 | 95.05 | 96.67 | 91.88 | 94.27 |
| Lung | 94.79 | 94.79 | 97.65 | 91.22 | 94.44 |
| Prostate | 91.96 | 92.09 | 94.03 | 86.70 | 90.36 |
| Average | 90.78 | 90.99 | 93.40 | 78.04 | 85.71 |

## Comparison of KARL to RL

In the following subsections, we present results for significance tests on whether KARL improves the Completeness, as well as, Consistency on the baseline model, RL. For ease of understanding, we depict the comparison as diagrams in this subsection and report the statistics on Completeness and Consistency for KARL and RL in Additional Files 2 and 3, respectively.

 doi:10.20944/preprints201908.0100.v1



*Comparison based on Completeness*

Figures 1a-c represent the difference in Completeness between KARL and RL. Figure 1a specifically represents the difference in maximum Coverage between KARL and baseline RL for normal training samples on all datasets. The 'Wins:Ties:Loses' ratio over that of the baseline was 12:13:0. Thus, Coverage on Control training samples of KARL was at least that of the baseline. That is, the difference is either zero or more. This observation was not too surprising as the additional background knowledge incorporated into KARL allows it to at least cover more training samples. Meanwhile, at a significance level of $\alpha = 0.05$, this gain in Coverage over Control samples was statistically significant ($p = 0.006069$ and $p = 0.001651$, using paired t-test and Wilcoxon signed-rank test, respectively).

Furthermore, Figure 1b shows the difference in maximum Coverage between KARL and baseline RL over the tumor training samples. Like the normal samples, the difference in Coverage for Cases is zero or more for KARL. The 'Wins:Ties:Loses' ratio as compared to the baseline was 9:16:0. The distribution of the 9 gains were 3, 2, 2, 1, 1 for prostate, brain, colon, lung and breast cancer, respectively. As expected, the general difference of Coverage between KARL and RL over the tumor examples was significant ($p = 0.01269$ and $p = 0.009152$). In general, the KARL method statistically significantly improves the baseline Coverage on both normal and tumor training samples.

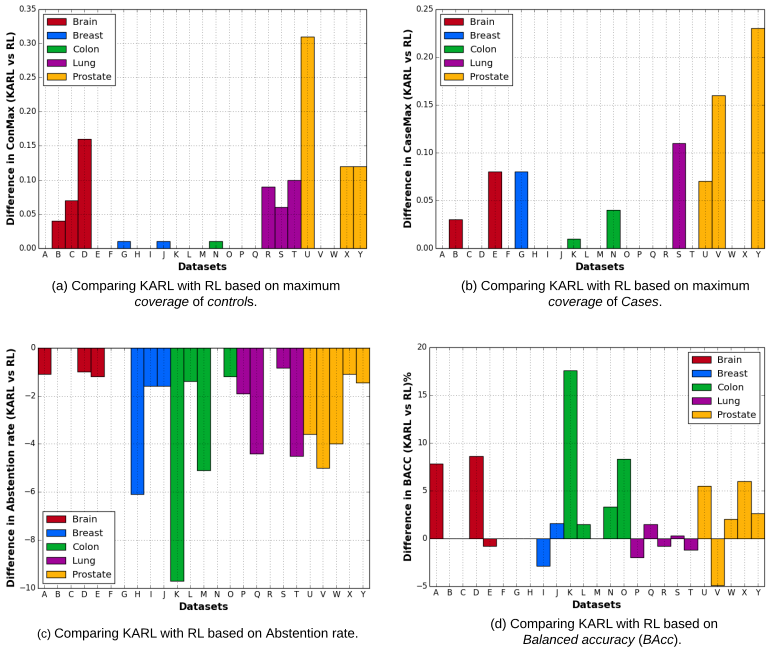
Figure 1c represents the difference in Abstention rate between KARL and RL models on the entire test datasets. As seen, KARL method reduced the Abstention rate in comparison to the baseline on most datasets, particularly, colon, prostate, and lung cancer sets. Its overall 'Wins:Ties:Loses' ratio over the Abstention rate on the baseline were 19:6:0. That is, for every dataset, its Abstention rate was as good as or better than that of baseline. This phenomenon was not surprising as a similar observation was made over Coverage. What is more, at a significance level of $\alpha = 0.05$, the reduction in Abstention rate by KARL was statistically significant ($p = 0.0001025$ and $p = 0.000143$). In general, KARL method statistically significantly improves baseline Abstention rate on both normal and tumor test examples.

*Comparison based on Consistency*

Figure 1d displays the difference in Consistency between KARL and RL on all datasets. The performance metric used here is BAcc. The 'Wins:Ties:Loses' ratio for our 25 datasets between KARL and RL was 13:6:6. In general, KARL made gains on the brain, colon, and prostate cancer sets; the gains were more pronounced in the latter two. In addition, using a paired t-test and a Wilcoxon signed-rank test, on all datasets, at a significance level of $\alpha = 0.05$, the classification performance of KARL, using BAcc, was significantly ($p = 0.0287$ and $p = 0.03624$, respectively) better than baseline RL. The spikes seen within the brain and colon cancer sets, for both methods, were largely due to gains in Specificity. For prostate cancer, however, a mix of gains caused the spikes observed from KARL in both Sensitivity and Specificity. Thus, gains in performance might depend on the nature of the dataset and the characteristics of the background knowledge that augments its modeling and interpretation. In general, the classification performance of KARL, using BAcc, statistically significantly improves baseline RL for the test data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1






(a) Comparing KARL with RL based on maximum *coverage* of *controls*.

(b) Comparing KARL with RL based on maximum *coverage* of *Cases*.

(c) Comparing KARL with RL based on Abstention rate.

(d) Comparing KARL with RL based on *Balanced accuracy* (*BAcc*).

Figure 1: **Comparing the KARL and RL models learned on cancer datasets.** Figure (a) compares models based on maximum Coverage of normal examples. The maximum Coverage of the tumor examples for both models are also compared in Figure (b). Models are compared based on their propensity to abstain from making predictions in Figure (c). Finally, Figure (d) shows the comparison of the models'performance based on their Balanced Accuracy (BAcc) measures.

## Domain and Data Involvement in the KARL Model

Our choice of transferring the knowledge of domain in the form of prior rules to KARL, enables us to track the contribution of sources in the learning process. Using the simple trick of tagging rules with their source of origin, Pr for domain and Nr for data, we can differentiate the rules based on their source of origin in the final model. We designed and used indicators, such as DataR and DataV to keep track of the total number of rules and variables in a KARL model, respectively. Other indicators include: DomainR and DomainV, for the total number of prior rules and their variables in the model, respectively. Then, DomainR and DomainV can be used for estimating the influence of the domain in the final and still-evolving models' rules and variables; similarly, DataR and DataV, can be used to estimate the involvement of the data. Accordingly, we designed $\frac{DomainR}{DomainR+DataR}$ as the domain-involvement ratio and $\frac{DataR}{DomainR+DataR}$ as the data-involvement ratio.

Table 6 presents the statistics over domain and data indicators, the involvement ratios for our cancer datasets. Studying the involvement ratios, we conclude that both domain and data contribute in building final model for almost all 25 datasets. We realized relatively noticeable range of values for domain-involvement ratio, [0.28, 1.0], except the zero value for GEO10780. Moreover, a considerable involvement of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1



Table 6: Involvement of domain & data in the KARL model. Domain indicators cover DomainR/DomainV which is the No. of priors rules/variables in the model. Data indicators cover DataR/DataV which is the No. of rules/variables in the model, generated from dataset. Involvement ratios contain domain-involvement and data-involvement ratios.

| | | Domain Indicators | | Data Indicators | | Involvement Ratios | |
|---|---|---|---|---|---|---|---|
| Disease | Dataset | DomainR | DomainV | DataR | DataV | Domain | Data |
| Brain Cancer | GEO16011 | 9 | 8 | 21 | 21 | 0.300 | 0.700 |
| | GEO1993 | 11 | 11 | 10 | 10 | 0.524 | 0.476 |
| | GEO4271 | 13 | 13 | 28 | 28 | 0.317 | 0.683 |
| | GEO4290 | 11 | 11 | 21 | 21 | 0.344 | 0.656 |
| | GEO4412 | 24 | 24 | 12 | 12 | 0.667 | 0.333 |
| Breast Cancer | GEO10780 | 0 | 0 | 22 | 21 | 0.000 | 1.000 |
| | GEO15852 | 7 | 7 | 18 | 18 | 0.280 | 0.720 |
| | GEO29431 | 2 | 2 | 2 | 2 | 0.500 | 0.500 |
| | GEO42568 | 5 | 5 | 1 | 1 | 0.833 | 0.167 |
| | GEO7904 | 9 | 9 | 12 | 12 | 0.429 | 0.571 |
| Colon Cancer | GEO10715 | 5 | 5 | 5 | 5 | 0.500 | 0.500 |
| | GEO20916 | 2 | 1 | 2 | 1 | 0.500 | 0.500 |
| | GEO23878 | 2 | 2 | 2 | 1 | 0.500 | 0.500 |
| | GEO24514 | 5 | 4 | 3 | 3 | 0.625 | 0.375 |
| | GEO9348 | 2 | 2 | 2 | 2 | 0.500 | 0.500 |
| Lung Cancer | GEO10072 | 2 | 1 | 0 | 0 | 1.000 | 0.000 |
| | GEO18842 | 2 | 1 | 2 | 2 | 0.500 | 0.500 |
| | GEO19188 | 5 | 4 | 4 | 4 | 0.556 | 0.444 |
| | GEO19804 | 15 | 13 | 9 | 9 | 0.625 | 0.375 |
| | GEO7670 | 6 | 6 | 6 | 6 | 0.500 | 0.500 |
| Prostate Cancer | GEO17951 | 13 | 13 | 19 | 19 | 0.406 | 0.594 |
| | GEO32448 | 13 | 13 | 9 | 9 | 0.591 | 0.409 |
| | GEO46602 | 2 | 1 | 0 | 0 | 1.000 | 0.000 |
| | GEO6956 | 9 | 9 | 11 | 1 | 0.450 | 0.550 |
| | GEO82188 | 17 | 13 | 21 | 21 | 0.447 | 0.553 |

the data can be noticed from the range of values for data-involvement ratio, [0.167, 1.0], except the zero values for GEO10072 and GEO46602.

We obtained more detailed insight on how KARL manages the use of data and domain in its learning process by studying these exceptions as mentioned above. All of these exceptions have acceptable values on Robustness, herein calculated as [Abstention rate, BAcc] for comparison purposes. As can be seen in Additional Files 2 and 3, Robustness is [1.1, 71.2] for GEO10780, [0.0, 97.9] for GEO10072, and [0.0, 94.4] for GEO46602. Abstention rate refers to the average percentage of test samples not assigned a class and BAcc refers to the average Balanced Accuracy over 10-fold cross validation, and these three datasets show acceptable Robustness and therefore useful for examining KARL's behavior. The KARL model on GEO10780 is an example of cases when domain background knowledge is not helpful for modeling the task on the target dataset. In this case, KARL recognizes and omits such rules in the pruning step of generating evidence-based prior rules which results in low values of DomainR, such as zero for GEO10780. Such datasets are interesting for further study, as they may have been obtained from a sample of a distinct population, with its own specific patterns not obeying the general domain patterns extracted from the background knowledge. On the other hand, KARL models on GEO10072 and GEO46602 datasets are examples of cases in which datasets cannot help learning Robust models. This can be the result of noise or inherent sample heterogeneity in the datasets. In these cases, the rules generated from these datasets are excluded for not reaching the user-defined constraints in RL learning process, resulting in zero values for DataR. However, these two datasets are modeled Robustly by rules borrowed from domain knowledge.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1



### Analysis of Patterns

In this subsection, we analyze whether intuitive and distinct rule patterns could be discovered across multiple KARL models for the related studies (e.g., same cancer type). Table 7 displays a snippet of these patterns discovered with KARL on brain, lung, and prostate cancers, across several datasets within the same disease type. We have used the statistics of each rule in addition to our desired thresholds for CF and Coverage of the rules, to select the patterns. For ease of presentation, we have covered a general and truncated version of these rules in Table 7. General in the sense that we use general placeholders, High and Low, for the discretized values of the rules' variables, and truncated in the sense that we have removed the statistics of rules. We provide the patterns learned for the breast and colon cancers, and the original version of the rules that are reported in Table 7, in Additional Files 4 and 5, respectively.

Table 7: A snippet of rule patterns discovered for brain, lung, and prostate cancers. Each row contains rule pattern information, including the learned models of the rule, its identifier, and its source of origin. Pr denotes that the pattern originates from prior knowledge, and Nr denotes new rules. $X$ is also used to build a nomenclature for variables of a gene family appears in the rule patterns.

| ID (Brain) | Rule ID | Source | Patterns |
|---|---|---|---|
| A, B, D, E | 01 | Pr, Nr | IF (COLXXX = Low) THEN (Class = Control) |
| | 02 | Pr, Nr | IF (COL6A3 = High) THEN (Class = Case) |
| A, B, D, E | 03 | Pr | IF (VEGFA = Low) THEN (Class = Control) |
| | 04 | Pr | IF (VEGFA = High) THEN (Class = Case) |
| | 05 | Pr | IF (LDHA = Low) THEN (Class = Control) |
| | 06 | Nr | IF (LDHA = High) THEN (Class = Case) |
| A, B, C, D | 07 | Pr | IF (IGFBPX = Low) THEN (Class = Control) |
| | 08 | Pr | IF (IGFBP2 = High) THEN (Class = Case) |
| B, E | 09 | Pr | IF (SERPINXX = Low) THEN (Class = Control) |
| A | 10 | Pr | IF (SERPINH1 = High) THEN (Class = Case) |

| ID (Lung) | Rule ID | Source | Patterns |
|---|---|---|---|
| P | 11 | Pr, Nr | IF (EDNRB = High) THEN (Class = Control) |
| | 12 | Nr | IF (EDNRB = Low) THEN (Class = Case) |
| | 13 | Pr | IF (PECAM1 = High) THEN (Class = Control) |
| | 14 | Pr | IF (PECAM1 = Low) THEN (Class = Case) |
| Q, R | 15 | Pr, Nr | IF (PLK4 = Low) THEN (Class = Control) |
| | 16 | Pr | IF (PLK4 = High) THEN (Class = Case) |
| | 17 | Pr, Nr | IF (AQP4 = High) THEN (Class = Control) |
| | 18 | Pr, Nr | IF (AQPX = Low) THEN (Class = Case) |
| S | 19 | Pr | IF (AGER = High) THEN (Class = Control) |
| | 20 | Pr | IF (AGER = Low) THEN (Class = Case) |
| | 21 | Pr | IF (CDH3 = Low) THEN (Class = Control) |
| | 22 | Pr | IF (CDH3 = High) THEN (Class = Case) |
| Q, S | 23 | Nr | IF (CCNB1 = Low) THEN (Class = Control) |
| | 24 | Pr | IF (CCNB1 = High) THEN (Class = Case) |

| ID (Prostate) | Rule ID | Source | Patterns |
|---|---|---|---|
| U, W, Y | 25 | Pr, Nr | IF (HPN = Low) THEN (Class = Control) |
| | 26 | Pr, Nr | IF (HPN = High) THEN (Class = Case) |
| Y | 27 | Pr | IF (TRPM4 = Low) THEN (Class = Control) |
| | 28 | Pr | IF (TRPM4 = High) THEN (Class = Case) |
| V, W | 29 | Pr, Nr | IF (CYP3A5 = High) THEN (Class = Control) |
| | 30 | Pr, Nr | IF (CYP3A5 = Low) THEN (Class = Case) |
| V, Y | 31 | Pr | IF (ID4 = High) THEN (Class = Control) |
| | 32 | Pr | IF (ID4 = Low) THEN (Class = Case) |

Studying the rules in Table 7, we can see that almost all patterns appear with their complementary rule, wherein the Condition for one rule leads to prediction of Case

 doi:10.20944/preprints201908.0100.v1



for example, and the negation of that Condition leading to prediction of Control (or not Case). It is also interesting that for some models, the complement of a rule was obtained from a combination of prior rule with that learned from target data. LDHA, for instance, appears in the Condition part of two complementary rules - while one is a retained prior rule (rule 05), the other is a new one (rule 06). These two features of our resulted patterns not only show that our strategy of directly involving prior rules in the KARL final model successfully achieves complementary rules, but also denote that both domain and data are contributing in achieving such patterns for our cancer prediction task.

A high level of consistency can also be noticed in the patterns of Table 7. Listed in the first column of the table, we can see that the polarity (i.e., whether the expression intensity of the gene is High or Low) for each model predictor variable was consistent across all models within which it occurred. As another interesting feature, we can see that a noticeable number of these patterns contain variables that belong to the same gene families. The uniqueness about such variables is that their polarity was almost consistent when they occurred in rules of the learned models (an example can be found in Table 8, snippets from brain cancer models.) For simplifying our results in Table 7, we have grouped all these variables as one variable and attributed a new name (nomenclature) to it. We designed the nomenclatures as a part of the gene family name followed by a few $X$s. For example, we represent those variables of the collagen family that are depicted in rules of Table 8, by using COLXXX as the nomenclature, as depicted in the simplified rule in Table 7.

Table 8: Example of unique rules from gene families for brain cancer.

| Source | Pattern |
|---|---|
| GEO1993 | IF (COL4A2=Low) THEN (Class=Control) |
| GEO4412 | IF (COL1A1=Low) THEN (Class=Control) |
| GEO16011 | IF (COL6A1=Low) THEN (Class=Control) |

## Discussion

The unique attributes of the patterns as elucidated above can be particularly useful for screening, diagnosis, and prognosis of same types of cancer. Though they require further and in-depth verification studies from domain experts, information contained in majority of them can be verified from literature. The members of collagen family of genes that have been discovered from the brain cancer (see Table 8), for instance, have been implicated in glioblastoma tumorigenesis; diffuse invasion of tumor cells into brain tissue typifies the advancement of tumor growth in some type of brain cancer, like glioblastoma [43, 44]. Senner et al. [43] found that the expression of collagen XVI was up-regulated in glioblastoma and it promotes tumor cell adhesion. Meanwhile, studies done by Bauer et al. [44] also found out that the inhibition of collagen XVI expression reduces glioma cell invasiveness. The first two rules depicted in Table 7 show that KARL also extracted such knowledge.

Moreover, AGER, also known as RAGE, is a member of the immunoglobulin super-family, and a multi-functional receptor with multiple ligands that have been found to play leading roles in diseases like arthritis, diabetes, and Alzeimer 's [45,46]. Evidence from recent studies indicate that this receptor likely plays an important

 doi:10.20944/preprints201908.0100.v1



role in cancer, particularly, its ability to lead cancer cell proliferation, invasion, and survival [45, 47, 48]. AGER, as well as several of its isoforms, is highly expressed in normal lung. However, and unlike other cancers, it is characterized by low expressions in human lung carcinomas [49, 50]. Thus, a down-regulated AGER would most likely be associated with a late stage of cancer, and therefore suggests it may function as a tumor suppressor for lung cancer. What is more, the general hypothesis, as elicited from literature above, that a highly expressed AGER implies normal, while the converse is true, was duly captured by KARL as rules (19) and (20) in Table 7.

Last, let us consider HPN (Hepsin), another biomarker and a variable that was discovered from the prostate cancer models (see Table 7). It contains a transmembrane serine protease, which may be in several cellular functions such as cell morphology and blood coagulation [51]. Several studies have identified Hepsin as one of the most up-regulated genes in prostate cancer [51, 52]. Observe that it was predominant and pervasive among most of the prostate cancer models. Similarly, the general notion in literature as regards correlation of its state of expression to prostate cancer progression was transferred into the KARL model, rules (25) and (26) in Table 7. Combining the evidence revealed from the examples above with other biologically unconfirmed/unverified patterns discovered by KARL turn it into a potent tool for cancer diagnosis and screening.

## Conclusions

In this paper, we develop and evaluate the novel KARL framework as an extension to rule learning by augmenting and transferring extant domain knowledge when modeling multiple types of cancer classification datasets to extract Robust underlying disease bio-mechanisms. Empirical results from our experiments highlight a couple of interesting features for the KARL's final model on our cancer datasets. First, we show that the knowledge augmented rule learning with KARL produces, on average, rule models that are more Robust classifiers than baseline RL without any background knowledge, using 25 publicly available gene expression datasets. This shows that our choice of domain knowledge for cancer prediction, in addition to, the transfer learning strategies designed for KARL improved the learning process over our 25 datasets. Second, we showed that KARL models yield biological patterns that underlie disease classification, using a combination of markers from both background knowledge and datasets. Third, we show that KARL provides the users with rule models which includes complementary classification rules. Fourth, we notice that KARL detected some variables which are from the same gene family and show consistent behavior by achieving consistent polarity and Class value for the rules which the family members appear in.

Though our preliminary empirical results suggest that the KARL framework is sound, we have identified potential limitations and several avenues for future work. The first limitation is on the type of variables extracted from the background domain knowledge, wherein these variables must already exist within the dataset being modeled. The second limitation is on the domain, which should be rich enough to provide KARL with a reliable background domain knowledge. Future work could deal with the first limitation by borrowing domain adaptation methods for transferring data between knowledge sources with different feature representations. This

 doi:10.20944/preprints201908.0100.v1



will help the model to accept any other type of knowledge as the unit of transfer. Moreover, future work could focus on expanding and automating the process of knowledge extraction from background domain. Like the work proposed in [53], this knowledge could be extracted from knowledge bases with hierarchical structure which is supported by automatic methods for extracting lookup tables. In this era of precision medicine, the second limitation should be easily overcome as knowledge accumulates from the vast numbers of biomarker studies being performed to generate, store and analyze panomics data for diverse diseases. KARL can then be used to examine accumulating evidence in an incremental fashion to extract interpretable and robust biological patterns that underlie health and disease.

## Methods

KARL provides users a framework in which the subjective knowledge of domain can be *incorporated* and *studied* in the learning process of an interpretable rule-learning model on the objective knowledge, the datasets. KARL is designed to be mainly applied to the sparse high-dimensional datasets which are highly likely to yield models that are overfit to training data and therefore underperfom on the test data. In order to avoid generating such models, KARL provides a transfer learning framework, wherein it can *incorporate* an additional source of knowledge for the learning process, namely, the subjective domain knowledge. The domain knowledge is expected to be a more reliable source of transfer, compared to direct transfer of rules learned from another noisy dataset used as the choice of transfer in the earlier rule transfer frameworks, TRL and TRL-FM. Domain 's experts, literature, or knowledge bases can be used as the reliable sources of domain knowledge required for KARL.

The choice of selecting RL as the building block of KARL, in addition to, the transfer learning strategy for incorporating the knowledge of domain in the model learning process, has originated from the main goal of designing a framework that can visualize the interaction of both domain and data sources in the learning process. We have designed KARL to transfer the knowledge of domain in the form of prior rules, which can be easily tracked during the RL learning process on data. The KARL final model then would consist of two types of rule: the prior rules transferred from domain, in addition to, the new rules learned from data. This visualization property of KARL helps scientists to *study* their subjective knowledge of interest along with data, which may lead the scientists to verify their hypothesis over their domain knowledge of interest over a subgroup of data. In this section, we elaborate upon the KARL method designed for the cancer prediction task. In the KARL Framework subsection, we explain the KARL's learning process. We then expand on the visualization property of KARL in the KARL Rule Model subsection, using our examples of rules learned on the cancer datasets.

### KARL Framework

We have designed KARL framework in three main steps: (1) Generating Lookup Table, (2) Generating Prior Rules, and (3) Generating KARL Rule Model. In this section, we elaborate upon the KARL framework, and how it is adapted for our use toward modeling for cancer prediction from gene expression data, as also shown in Figure 2.

 doi:10.20944/preprints201908.0100.v1



*Generating Lookup Table*

In this step, the user should provide KARL with the knowledge of domain that should be structured in the form of a table of variables that also exists in the target dataset. The variables of this table, the lookup table, can be sourced from domain's experts, literature, knowledge bases, or other reliable databases. For our task of cancer prediction, we propose using the hierarchical domain knowledge, such as biological pathways, as our source of domain knowledge. As shown in Figure 2a, we derive the variables of the lookup table to be those gene expression variables that are extracted from the cancer dataset ($I_D$) and are associated with the general properties of cancer ($I_P$) through the cancer hierarchical background knowledge ($I_K$). The lookup table generation process for the cancer prediction task, in addition to its requirements are detailed below.

A domain can be characterized by actors, their interactions, and properties. KARL posits that a rule pattern is interesting if evidence can be drawn from desirable sources of the variables (actors) that represent the actionable aspects of the domain (properties). Thus, at the outset, it is essential for KARL to identify and delineate a domain, including its actors and properties. To this end, we have selected genetic actors—genes—and their properties for our domain of study, cancer. These gene actors are selected as the variables of the input gene-expression dataset, that undergoes discretization as a pre-processing step to bin the continuous valued measurements into discrete ranges, so as to reduce the search space of possible models that need to be generated. As the discretization method, we have used the Efficient Bayesian Discretization (EBD) [54]. EBD uses a Bayesian score, which has a lambda prior parameter, to discover the optimal number of bins automatically. For our experiments, we set the value for this parameter to 0.5, which is the default. In this manuscript, by datasets ($I_D$), we refer to their discretized version.

For a cell to progress into a tumor, it acquires a whole gamut of aberrant properties. While different cancer types may require different combinations of these properties, typical behaviors (or hallmarks) that underpin them can be categorized. Seminal work done by Hanahan and Weinberg [55,56] has suggested that an extensive catalog of cancer cell genotypes is a manifestation of six main capabilities that turn faulty to modify the physiology of the cell. These hallmarks can be merged further into three main properties. The first is the faulty control of the cell cycle (i.e., sustaining proliferative signaling and evading growth suppressors hallmarks); the second property is the faulty control of cell death (i.e., resisting cell death and enabling replicative immortality hallmarks); and the third property is the invasiveness and metastatic capabilities (i.e., inducing angiogenesis and activating invasion and metastasis hallmarks). Using KARL for cancer prediction, we have proposed these three hallmarks as the properties of domain ($I_P$).

Background domain knowledge ($I_K$) can be sourced from an expert, literature, or knowledge bases. The main idea and challenge in using background knowledge in KARL is, how to incorporate such knowledge sources and determine those actors of the domain functioning as the desired properties of the domain. For the cancer prediction domain, we selected the Ingenuity® Knowledge Base (IKB). The IKB contains evidence-based domain knowledge in the form of gene-interaction networks. This knowledge base can be mined and explored by using the Ingenuity® Pathway

 doi:10.20944/preprints201908.0100.v1



Analysis (IPA) tool. Given a list of genes, IPA creates and outputs molecular networks (algorithmically generated pathways) by using hypergeometric testing to map each gene to the information contained in the knowledge base [57]. In order to generate the lookup table, we gave as input to the IPA, the list of gene variables in the input dataset, to obtain the genes'corresponding functional networks from IKB. Then, these networks are ranked according to a $P_{score}$ = -$\log_{10}$(p-value), where the p-value is derived from a hypergeometric test. When a threshold of $P_{score} = \alpha$ is applied, all non-significant networks (i.e., $P_{score} < \alpha$) are removed. We set the threshold ($\alpha$) to 2.0. for this work. Any of the input genes that has significant network(s) with functionality of at least one of the hallmarks of cancer is then added to the lookup table. A part of a lookup table generated for the brain cancer datasets is depicted in Figure 2b.

*Generating Prior Rules*

The background domain knowledge should first be transformed into the form of IF-THEN propositional rules to fit into the KARL design. KARL uses these prior rules in two steps to augment the rule learning model (see following subsection for details). In this subsection, we mainly describe the process of generating this set of rules, depicted as Step 2 in Figure 2a. KARL's prior rule generation engine induces prior rules based on the input list of gene variables from the lookup table. This engine outputs single-variable prior rules that result from all possible combinations of gene and phenotype values. For example, assume that the gene expression variable, VEGFA, takes two values: High and Low, which are the placeholders for the discretized bins of values for each gene's expression.Also assume that the phenotype, Class, take two values: Case and Control. Then, the prior rules for VEGFA can be instantiated with a permutation of all values for it as a predictor of the phenotype. These rules are reported below and are also depicted in Figure 2b:

1. IF (VEGFA is High) THEN (Class = Case)
2. IF (VEGFA is High) THEN (Class = Control)
3. IF (VEGFA is Low) THEN (Class = Case)
4. IF (VEGFA is Low) THEN (Class = Control)

The output from this engine will be the combination of such prior rules generated for all variables of the lookup table. This set is then used for transfer learning to generate the KARL rule model, as described below.

*Generating KARL Rule Model*

The KARL rule model, as depicted in Step 3 of Figure 2a, is a combination of two rule sets: (1) Rule Model with Prior Rules, and (2) Evidence-based Prior Rules. KARL produces these two rule sets by using two transfer learning strategies that use the outcome from the previous step, that is the prior rules. The Rule Model with Prior Rules is the model learned by RL with Priors ($RL_P$) on the input dataset ($I_D$) and the user-defined criteria ($I_C$) for acceptable model quality specified as values for the set of parameters for RL's search engine. $RL_P$ is a modified version of RL in which prior rules are transferred to RL to be used as the initial general rule model that the learning process starts from. Evidence-based Prior Rules is the subset of prior rules pruned on the input dataset ($I_D$) in an evaluation process

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1



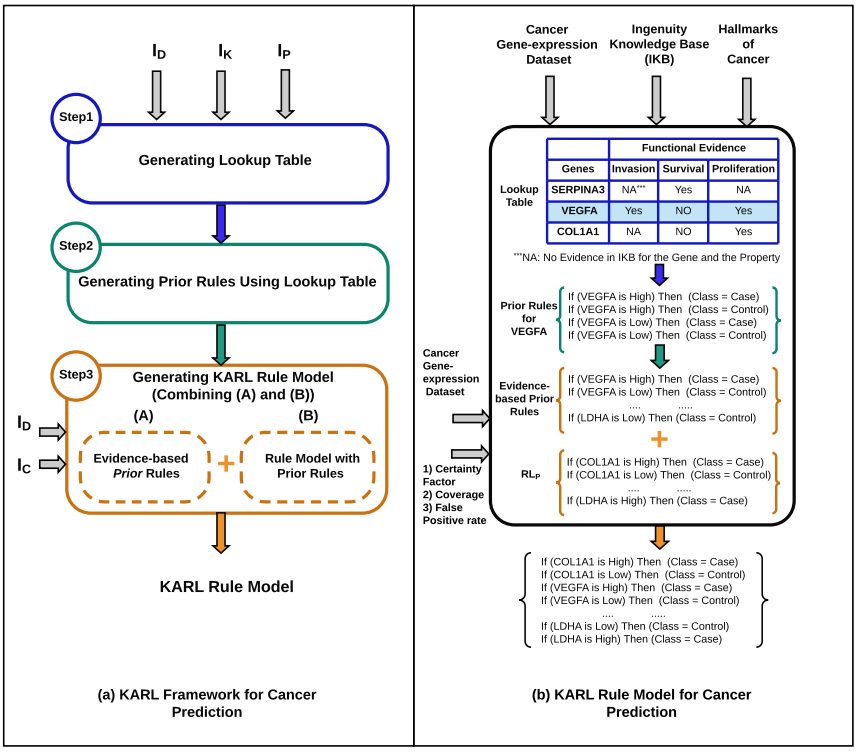



Figure 2: **Knowledge Augmented Rule Learning (KARL) framework.** Panel (a) depicts the KARL framework for cancer prediction within three main steps. Step 1 (generating lookup table) accepts three inputs: dataset ($I_D$), domain background knowledge ($I_K$), and properties of domain ($I_P$). Using this set of inputs in this step, the gene variables associated with the properties of domain through the background domain knowledge are selected and saved as a lookup table. The lookup table is then used in Step 2 (Generating Prior Rules), in order to generate the prior rules being used as the input of Step 3 (Generating KARL Rule Mode). In this step, the prior rules are used to generate two sets of rules: Rule Model with Prior Rules and Evidence-based Prior rules. Rule learner with priors ($RL_P$) uses the prior rules, as well as, the user-defined criteria ($I_C$), and datasets ($I_D$) in order to generate the Rule Model with Prior Rules. The subset of prior rules validated on dataset through the user-defined criteria is also selected as the Evidence-based Prior rules. The combination of these rule sets is then selected as the KARL rule model. Panel (b) depicts the outputs of steps in the KARL framework for cancer prediction task. For this task, the gene-expression data is used as ($I_D$), Ingenuity knowledge base (IKB) is selected as the domain background knowledge, and hallmarks of cancer are designed to be the properties of domains. Minimum Certainty Factor (CF), minimum Coverage, and maximum False Positive (FP) rate are also selected as the user-defined criteria ($I_C$). The lookup table and rule sets depicted in this panel are snippets of results from KARL model learned on the brain cancer datasets.

using user-defined criteria ($I_C$). This set has been designed as a source of domain knowledge getting transferred directly to the KARL's final model. This transfer learning strategy seeks to improve KARL by using hierarchical domain knowledge, such as biological pathways to learn more general rules that may not be extracted

 doi:10.20944/preprints201908.0100.v1



at the gene level due to the sparse data. The snippet of the Rule Model with Prior Rules, Evidence-based Prior Rules, and KARL model for brain cancer datasets are depicted in Figure 2b. In the following subsections, we elaborate upon the generation process for each of these sets.

***Rule Learners with and without Priors ($RL_P$ and RL)*** The $RL_P$ method produces a set of IF-THEN propositional rules extracted from both domain and data. Each rule has statistics associated with it, such as the Certainty Factor (CF) which indicates the degrees of belief or disbelief in that rule, based on the evidence seen in the data. Below, we depict an example of a rule, without its statistics, where High and Low are placeholders for the discretized bins of values for each gene's expression:

IF ((Gene1Exp is High) AND (Gene2Exp is Low)) THEN (Class = Case)

The pseudocode for the $RL_P$ algorithm is reported as the Additional File 1. This algorithm proceeds as a beam (priority queue-based) search of the model space using a general-to-specific approach to output a rule model. It starts with a general model, the priority-ordered prior rules generated from the lookup table, in decreasing order of each rule's CF value. This model is then used to initialize the priority-ordered queue used in $RL_P$. The size of the queue, in addition to, the priority metrics (user-defined criteria ($I_C$)) are parameters that should be specified by the user. There are several choices that users can provide as these criteria, from which we have selected the minimum Certainty Factor (CF), minimum Coverage (covered in more detail in Evaluation Metrics subsection), and maximum False Positive (FP) rate.

The rules satisfying these criteria are called *good* rules in the pseudocode. As the next step, $RL_P$ starts expanding the rule model by adding conjuncts (i.e., variable-value pairs) to the the Condition of rules, in order to specialize the existing rules; or learning new single-variable rules from the dataset. The queue is then replaced with a new priority-ordered queue filled with new generated rules and the former queue in order of decreasing CF value up to the maximum size of the queue. The $RL_P$ algorithm stops when the rule model stays unchanged. This is the resulting Rule Model with Prior Rules and is one part of the final KARL model.

As mentioned earlier, RL differs from $RL_P$ in its selection of the initial general rule set (Additional File 1, Line 5). RL starts with a random general model that contains a set of rules selected from the space of all possible one-variable Condition generated from all discretized variable values and Class labels. This set is a priority-ordered queue of rules which is being expanded by the same strategy as $RL_P$ to result in the RL model.

***Evidence-based Prior Rules*** Evidence-based Prior Rules is the subset of prior rules that satisfies the user-defined criteria ($I_C$) when using the input dataset $I_D$. The set of prior rules generated from the lookup table contains pairs of single-variable contradictory rules. The contradictory pairs contain rules with similar Conditions which is leading to different predictions. A pruning step is used to determine how to remove these contradictions from the prior rules generated. The misleading rule in the contradictory pair is recognizable by its low performance or Coverage on the data. We use this behavior of the misleading rules as the basis of our pruning method. The user-defined criteria that is used in the $RL_P$ method is capable

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1



of detecting such behavior. These criteria are a minimum Certainty Factor value, minimum Coverage, and maximum False Positive rate. This pruned subset of the prior rules is what we refer to as the Evidence-based Prior Rules. The combination of Rule Model with Prior Rules and Evidence-based Prior Rules is then selected as the final model for KARL, that is used for prediction on unseen test samples. In our experiments, we have used 0.8 for Positive Predictive Value (PPV) as the CF, 4 for the minimum Coverage, and 0.1 for the maximum False Positive rate, as the set of values for the user-defined criteria.

### KARL Rule Model

In Figure 2b, we depict a snippet of a modified version of KARL rule model for brain cancer. In this section, we elaborate upon the original format of rules generated by frameworks using RL or $RL_P$. We also explain the interpretation of the discovered patterns - the rules of the KARL model. We also show how rules can be tracked during the KARL learning process, in order to estimate and visualize the interaction of their sources of origin in the final model.

The following rule is the original rule extracted from KARL model on the brain cancer datasets, which also appeared in the final KARL model depicted in Figure 2b:

Pr. IF (VEGFA = -inf..8.68) THEN (@Class = Control)
CF=0.928, TP=13, FP=1, Pos=19, Neg=39

All the rules in their original format appear with the discretized bins of values for the variables of their Condition. An example of such bin is [-inf..8.68], which is assigned to the VEGFA variable in our example rule. This rule is regarded as a classifier which predicts Control Class for any sample with VEGFA less than or equal to 8.68. Each rule of the model is also accompanied with statistics, which evaluate the rule's Robustness. These statistics are collected based on the evidence seen in the data, and consist of Certainty Factor (CF), True Positive (TP), False Positive (FP), and number of Positive and Negative samples (Pos and Neg). In a classification task on a set of samples, each rule either abstains from predicting those samples that cannot satisfy the rule's Condition, or it always predict the rest of samples with its Class value. The supporting statistics are then designed to convey evaluation information on the rule's prediction, based on its Class value. Accordingly, the positive Class used in providing the rule's statistics are determined specifically for each individual rule based on the value represented in its right-hand-side. For example, the positive Class is Control for our rule shown above that is learned on the brain cancer dataset, while the positive Class is Case for the cancer prediction task. The statistics for a rule are then collected based on its prediction and positive and negative instances (Pos and Neg, respectively). These statistics are further used for evaluating the rule's Robustness, estimated over rule's Consistency and Completeness. While the CF is set to PPV and is a measure of rule's Consistency, the Rule's Completeness is measured as Coverage, which is calculated using the rest of the statistics. Coverage is defined as the fraction of the test samples covered (classified) by the rule, (TP + FP)/(Pos + Neg). The following interpretation can then be added to our rule, based on its statistics: the rule is a predictor of the Control class, which have covered (classified) 14 (1 + 13) samples over the total number of 58 (39 + 19) samples and resulted in 92.8% PPV prediction performance for the covered samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2019 doi:10.20944/preprints201908.0100.v1



KARL marks its rules with labels indicating the rule's source of origin: domain or data. It uses Pr for tagging the rules that originate from domain, prior rules, while it tags the rules that originate from the data, new rules, with Nr. For instance, our example rule is a prior rule as it has the Pr tag. This rule's simple trick of marking rules, provides the functionality to visualize the transfer and accordingly the interaction of domain and data in its final model. As an example of such interactions, we can determine the complementary rules as mentioned in the Analysis of Patterns section. We also believe that the involvement of a source in the final model can be estimated as the fraction of the KARL model's rules that originates from that source. Accordingly, we have designed the two following measures: domain-involvement ratio : $\frac{DomainR}{DomainR+DataR}$ and data-involvement ratio: $\frac{DataR}{DomainR+DataR}$, in which DomainR and DataR are the number of the rules tagged with Pr and Nr, respectively. We have also reported these estimations on our cancer datasets in the Domain and Data Involvement in the KARL Model section.

Any of the rule models learned by either RL, $RL_P$, or KARL is actually a set of rules as our example rule. These models use the same strategy for predicting the Class value of the test samples. For each unseen test sample, zero or more rules of the model could fire (i.e., the Condition part of the rule matches the corresponding marker/variable values for the test instance). Samples with zero fired rules are reported as Abstained samples, and are not given Class prediction by the model. In the case of conflicting predictions from two or more fired rules, different strategies are adopted, including: first matching rule, equal voting, and weighted voting. In the first matching rule method, the Class value of the rule with the highest CF and/or Coverage measures, is selected as the predicted Class. In the equal voting strategy, every matching rule contributes a single vote for its Class and the prediction with the majority votes wins. For weighted voting, each matching rule votes with a weight of its CF; the Class with the highest summed CF and/or Coverage wins. As our inference method for RL and KARL, we used weighted voting for dealing with conflict resolution.

**Additional Files**

**Additional File 1:** Rule learner with priors ($RL_P$) algorithm.
This file is the figure of the pseudocode for the $RL_P$ algorithm along with comments for each line of it. (PNG file 94 KB)

**Additional File 2:** Average Completeness on 10-Fold cross-validation for KARL vs RL.
This file contains a table comparing KARL with RL on 25 cancer datasets used in the paper. The comparison is based on measures including: ConMdn = median Coverage for Controls, CaseMin = minimum Coverage for Cases, CaseMax = maximum Coverage for Cases, CaseMdn = median Coverage for Cases, and AbsRate = %test samples not assigned a Class. (PDF file 291 KB)

**Additional File 3:** Average Consistency percentage on 10-fold cross-validation for KARL vs RL.
This file contains a table comparing KARL with RL on 25 cancer datasets used in the manuscript. The comparison is based on measures including: AccAb = Accuracy including Abstentions, Acc = Accuracy, SN = Sensitivity, SP = Specificity, BAcc = Balanced Accuracy. (PDF file 49 KB)

**Additional File 4:** A snippet of patterns discovered for breast and colon cancers.
This file contains a table of rule pattern information for breast and colon cancers, including the learned models of the rule, the rule identifier, the source of rule origin, and the rule patterns. (PDF file 53 KB)

**Additional File 5:** Rules of Table 7 in their original representation.
This file contains the original version of rules covered in Table 7, i.e., with their statistics, discretized values for variables, and labels of their source of origin. Each row of the table contains the ID, the dataset, and the original pattern of the corresponding rule in Table 7. (PDF file 79 KB)

**Abbreviations**
KARL:Knowledge Augmented Rule Learning; BAcc: Balanced Accuracy; IKB: Ingenuity Knowledge Base; IPA: Ingenuity Pathway Analysis; ANOVA: Analysis of Variance; GSEA: Gene Set Enrichment Analysis; ANN: Artificial Neural Networks; SVMs: Support Vector Machines; TRL: Transfer Rule Learning; RL: Rule LEarner; AUC: Area Under ROC Curve; EBD: Efficient Bayesian Discretization; CF: Certainty Factor; PPV: Positive Predictive Value; GEO: Gene Expression Omnibus; SN: Sensitivity; SP: Specificity; Acc: Accuracy; AccAb: Accuracy including Abstention; LR: Logistic Regression; NB: Naive Bayes; RF: Random Forests; ConMin: minimum coverage for Controls; ConMax: maximum; coverage for Controls; ConMdn: median coverage for Controls; CaseMin: minimum coverage for Cases; CaseMax: maximum coverage for Cases; CaseMdn: median coverage for Cases; AbsRate: %test samples not assigned a class; RLP: RL with Priors; FP: False Positive

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Availability of data and materials**
The public data analyzed in this study is downloaded from Gene Expression Omnibus (GEO) repository. All the references to this data were reported in Table 1 of the manuscript too. We have provided KARL along with sample input and output and README files/manual at https://github.com/Mahbaneh/KARL.

**Competing interests**
The authors declare that they have no competing interests.

**Funding**
The authors gratefully acknowledge grants from the National Institute of General Medical Sciences of the National Institutes of Health (R01GM100387) and the National Library of Medicine Training Grant (5T15LM007059-26). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Authors'contributions**
HAO designed and implemented the KARL algorithm, performed the experiments. MET verified the algorithm, modified the framework, added some experiments, and drafted the manuscript. VG oversaw the study, developed the initial concept of discovering robust biological patterns from diverse sources using transfer learning, provided HAO and MET with the RL code extensions developed in her laboratory, and helped with experimental design. All authors have contributed to the revisions of the manuscript and approve of the final version.

**Acknowledgements**
The authors also thank Jeya B. Balasubramanian from the PRoBE laboratory in the Department of Biomedical Informatics, for helping run some additional experiments efficiently in order to address reviewer comments. We also thank Jeya Balaji Balasubramanian for helping us conducting some experiments.

**Author details**
[1]Department of Biomedical Informatics, University of Pittsburgh, 5607 Baum Bld, Pittsburgh, PA 15206, USA. [2]Department of Biomedical Engineering, University of Ghana, Accra, Ghana. [3]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA USA. [4]Department of Computational & Systems Biology, Pittsburgh, PA USA. [5]Department of Bioengineering, Pittsburgh, PA USA.

**References**
1. Walker, M.S., Hughes, T.A.: Messenger rna expression profiling using dna microarray technology: Diagnostic tool, scientific analysis or un-interpretable data? International journal of molecular medicine **21**(1), 13–17 (2008)
2. Shendure, J., Ji, H.: Next-generation dna sequencing. Nature biotechnology **26**(10), 1135–1145 (2008)
3. Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nature reviews genetics **10**(1), 57–63 (2009)
4. Ozsolak, F., Milos, P.M.: Rna sequencing: advances, challenges and opportunities. Nature reviews genetics **12**(2), 87–98 (2011)
5. Malone, J.H., Oliver, B.: Microarrays, deep sequencing and the true measure of the transcriptome. BMC biology **9**(1), 34 (2011)
6. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y.: Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research **18**(9), 1509–1517 (2008)
7. Schulze, A., Downward, J.: Navigating gene expression using microarrays—a technology review. Nature cell biology **3**(8), 190–195 (2001)
8. Slonim, D.K., Yanai, I.: Getting started in gene expression microarray analysis. PLoS computational biology **5**(10), 1000543 (2009)
9. Ganchev, P., Malehorn, D., Bigbee, W.L., Gopalakrishnan, V.: Transfer learning of classification rules for biomarker discovery and verification from molecular profiling studies. Journal of biomedical informatics **44**, 17–23 (2011)
10. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big Data **3**(1), 9 (2016)
11. Clearwater, S.H., Provost, F.J.: Rl4: A tool for knowledge-based induction. In: Tools for Artificial Intelligence, 1990., Proceedings of the 2nd International IEEE Conference On, pp. 24–30 (1990). IEEE
12. Ogoe, H.A., Visweswaran, S., Lu, X., Gopalakrishnan, V.: Knowledge transfer via classification rules using functional mapping for integrative modeling of gene expression data. BMC bioinformatics **16**(1), 226 (2015)

13. Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic acids research **30**(1), 207–210 (2002)
14. Gravendeel, L.A., Kouwenhoven, M.C., Gevaert, O., de Rooi, J.J., Stubbs, A.P., Duijm, J.E., Daemen, A., Bleeker, F.E., Bralten, L.B., Kloosterhof, N.K., *et al.*: Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. Cancer research **69**(23), 9065–9072 (2009)
15. Petalidis, L.P., Oulas, A., Backlund, M., Wayland, M.T., Liu, L., Plant, K., Happerfield, L., Freeman, T.C., Poirazi, P., Collins, V.P.: Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. Molecular cancer therapeutics **7**(5), 1013–1024 (2008)
16. Phillips, H.S., Kharbanda, S., Chen, R., Forrest, W.F., Soriano, R.H., Wu, T.D., Misra, A., Nigro, J.M., Colman, H., Soroceanu, L., Williams, P.M., Modrusan, Z., Feuerstein, B.G., Aldape, K.: Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer cell **9**(3), 157–173 (2006)
17. Sun, L., Hui, A.-M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R., *et al.*: Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer cell **9**(4), 287–300 (2006)
18. Freije, W.A., Castro-Vargas, F.E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L.M., Mischel, P.S., Nelson, S.F.: Gene expression profiling of gliomas strongly predicts survival. Cancer research **64**(18), 6503–6510 (2004)
19. Zhang, Z., Chen, D., Fenstermacher, D.A.: Integrated analysis of independent gene expression microarray datasets improves the predictability of breast cancer outcome. BMC genomics **8**(1), 331 (2007)
20. Ni, I.B.P., Zakaria, Z., Muhammad, R., Abdullah, N., Ibrahim, N., Emran, N.A., Abdullah, N.H., Hussain, S.N.A.S.: Gene expression patterns distinguish breast carcinomas from normal breast tissues: the malaysian context. Pathology-Research and Practice **206**(4), 223–228 (2010)
21. Lopez, F.-J., Cuadros, M., Cano, C., Concha, A., Blanco, A.: Biomedical application of fuzzy association rules for identifying breast cancer biomarkers. Medical & biological engineering & computing **50**(9), 981–990 (2012)
22. Clarke, C., Madden, S.F., Doolan, P., Aherne, S.T., Joyce, H., O'Driscoll, L., Gallagher, W.M., Hennessy, B.T., Moriarty, M., Crown, J., *et al.*: Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. Carcinogenesis **34**(10), 2300–2308 (2013)
23. Richardson, A.L., Wang, Z.C., De Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J.D., Livingston, D.M., Ganesan, S.: X chromosomal abnormalities in basal-like human breast cancer. Cancer cell **9**(2), 121–132 (2006)
24. Galamb, O., Sipos, F., Solymosi, N., Spisák, S., Krenács, T., Tóth, K., Tulassay, Z., Molnár, B.: Diagnostic mrna expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with peripheral blood results. Cancer Epidemiology and Prevention Biomarkers **17**(10), 2835–2845 (2008)
25. Skrzypczak, M., Goryca, K., Rubel, T., Paziewska, A., Mikula, M., Jarosz, D., Pachlewski, J., Oledzki, J., Ostrowsk, J.: Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. PloS one **5**(10), 13091 (2010)
26. Uddin, S., Ahmed, M., Hussain, A., Abubaker, J., Al-Sanea, N., AbdulJabbar, A., Ashari, L.H., Alhomoud, S., Al-Dayel, F., Jehan, Z., *et al.*: Genome-wide expression analysis of middle eastern colorectal cancer reveals foxm1 as a novel target for cancer therapy. The American journal of pathology **178**(2), 537–547 (2011)
27. Alhopuro, P., Sammalkorpi, H., Niittymäki, I., Biström, M., Raitila, A., Saharinen, J., Nousiainen, K., Lehtonen, H.J., Heliövaara, E., Puhakka, J., *et al.*: Candidate driver genes in microsatellite-unstable colorectal cancer. International journal of cancer **130**(7), 1558–1566 (2012)
28. Hong, Y., Downey, T., Eu, K.W., Koh, P.K., Cheah, P.Y.: A 'metastasis-prone'signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. Clinical & experimental metastasis **27**(2), 83–90 (2010)
29. Landi, M.T., Dracheva, T., Rotunno, M., Figueroa, J.D., Liu, H., Dasgupta, A., Mann, F.E., Fukuoka, J., Hames, M., Bergen, A.W., *et al.*: Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PloS one **3**(2), 1651 (2008)
30. Sanchez-Palencia, A., Gomez-Morales, M., Gomez-Capilla, J.A., Pedraza, V., Boyero, L., Rosell, R., Fárez-Vidal, M.: Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. International journal of cancer **129**(2), 355–364 (2011)
31. Hou, J., Aerts, J., Den Hamer, B., Van Ijcken, W., Den Bakker, M., Riegman, P., van der Leest, C., van der Spek, P., Foekens, J.A., Hoogsteden, H.C., *et al.*: Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PloS one **5**(4), 10312 (2010)
32. Lu, T.-P., Tsai, M.-H., Lee, J.-M., Hsu, C.-P., Chen, P.-C., Lin, C.-W., Shih, J.-Y., Yang, P.-C., Hsiao, C.K., Lai, L.-C., et al.: Identification of a novel biomarker sema5a for non-small cell lung carcinoma in non-smoking women. Cancer Epidemiology and Prevention Biomarkers, 0332 (2010)
33. Jia, Z., Wang, Y., Sawyers, A., Yao, H., Rahmatpanah, F., Xia, X.-Q., Xu, Q., Pio, R., Turan, T., Koziol, J.A., *et al.*: Diagnosis of prostate cancer using differentially expressed genes in stroma. Cancer research **71**(7), 2476–2487 (2011)
34. Derosa, C., Furusato, B., Shaheduzzaman, S., Srikantan, V., Wang, Z., Chen, Y., Siefert, M., Ravindranath, L., Young, D., Nau, M., *et al.*: Elevated osteonectin/sparc expression in primary prostate cancer predicts metastatic progression. Prostate cancer and prostatic diseases **15**(2), 150 (2012)
35. Mortensen, M.M., Hoyer, S., Lynnerup, A.-S., Orntoft, T.F., Sorensen, K.D., Borre, M., Dyrskjot, L.: Expression profiling of prostate cancer tissue delineates genes associated with recurrence after prostatectomy. Scientific reports **5**, 16018 (2015)
36. Wallace, T.A., Prueitt, R.L., Yi, M., Howe, T.M., Gillespie, J.W., Yfantis, H.G., Stephens, R.M., Caporaso, N.E., Loffredo, C.A., Ambs, S.: Tumor immunobiological differences in prostate cancer between african-american and european-american men. Cancer research **68**(3), 927–936 (2008)
37. Wang, Y., Xia, X.-Q., Jia, Z., Sawyers, A., Yao, H., Wang-Rodriquez, J., Mercola, D., McClelland, M.: In silico estimates of tissue components in surgical samples based on expression profiling data. Cancer research **70**(16),

 doi:10.20944/preprints201908.0100.v1



6448–6455 (2010)
38. Danyluk, A.P., Provost, F.: Small disjuncts in action: learning to diagnose errors in the local loop of the telephone network. In: Proc. of Tenth International Conference on Machine Learning, pp. 81–88 (2014)
39. Ryberg, H., An, J., Darko, S., Lustgarten, J.L., Jaffa, M., Gopalakrishnan, V., Lacomis, D., Cudkowicz, M., Bowser, R.: Discovery and verification of amyotrophic lateral sclerosis biomarkers by proteomics. Muscle & nerve **42**(1), 104–111 (2010)
40. Ranganathan, S., Williams, E., Ganchev, P., Gopalakrishnan, V., Lacomis, D., Urbinelli, L., Newhall, K., Cudkowicz, M.E., Brown Jr, R.H., Bowser, R.: Proteomic profiling of cerebrospinal fluid identifies biomarkers for amyotrophic lateral sclerosis. Journal of neurochemistry **95**(5), 1461–1471 (2005)
41. Hennessy, D., Gopalakrishnan, V., Buchanan, B.G., Rosenberg, J.M., Subramanian, D.: Induction of rules for biological macromolecule crystallization. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, vol. 179, p. 187 (1994). AAAI Press
42. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, Massachusetts (2016)
43. Senner, V., Ratzinger, S., Mertsch, S., Grässel, S., Paulus, W.: Collagen xvi expression is upregulated in glioblastomas and promotes tumor cell adhesion. FEBS letters **582**(23-24), 3293–3300 (2008)
44. Bauer, R., Ratzinger, S., Wales, L., Bosserhoff, A., Senner, V., Grifka, J., Grässel, S.: Inhibition of collagen xvi expression reduces glioma cell invasiveness. Cellular Physiology and Biochemistry **27**(3-4), 217–226 (2011)
45. Logsdon, C.D., Fuentes, M.K., Huang, E.H., Arumugam, T.: Rage and rage ligands in cancer. Current molecular medicine **7**(8), 777–789 (2007)
46. Hudson, B.I., Carter, A.M., Harja, E., Kalea, A.Z., Arriero, M., Yang, H., Grant, P.J., Schmidt, A.M.: Identification, classification, and expression of rage gene splice variants. The FASEB Journal **22**(5), 1572–1580 (2008)
47. Malik, P., Chaudhry, N., Mittal, R., Mukherjee, T.K.: Role of receptor for advanced glycation end products in the complication and progression of various types of cancers. Biochimica et Biophysica Acta (BBA)-General Subjects **1850**(9), 1898–1904 (2015)
48. Sparvero, L.J., Asafu-Adjei, D., Kang, R., Tang, D., Amin, N., Im, J., Rutledge, R., Lin, B., Amoscato, A.A., Zeh, H.J., *et al.*: Rage (receptor for advanced glycation endproducts), rage ligands, and their role in cancer and inflammation. Journal of translational medicine **7**(1), 17 (2009)
49. Englert, J.M., Hanford, L.E., Kaminski, N., Tobolewski, J.M., Tan, R.J., Fattman, C.L., Ramsgaard, L., Richards, T.J., Loutaev, I., Nawroth, P.P., *et al.*: A role for the receptor for advanced glycation end products in idiopathic pulmonary fibrosis. The American journal of pathology **172**(3), 583–591 (2008)
50. Bartling, B., Hofmann, H.-S., Weigle, B., Silber, R.-E., Simm, A.: Down-regulation of the receptor for advanced glycation end-products (rage) supports non-small cell lung carcinoma. Carcinogenesis **26**(2), 293–301 (2005)
51. Klezovitch, O., Chevillet, J., Mirosevich, J., Roberts, R.L., Matusik, R.J., Vasioukhin, V.: Hepsin promotes prostate cancer progression and metastasis. Cancer cell **6**(2), 185–195 (2004)
52. Valkenburg, K.C., Hostetter, G., Williams, B.O.: Concurrent hepsin overexpression and adenomatous polyposis coli deletion causes invasive prostate carcinoma in mice. The Prostate **75**(14), 1579–1585 (2015)
53. Eshaghzadeh Torbati, M., Mitreva, M., Gopalakrishnan, V.: Application of taxonomic modeling to microbiota data mining for detection of helminth infection in global populations. Data **1**(3), 19 (2016)
54. Lustgarten, J.L., Visweswaran, S., Gopalakrishnan, V., Cooper, G.F.: Application of an efficient bayesian discretization method to biomedical data. BMC bioinformatics **12**(1), 309 (2011)
55. Hanahan, D., Weinberg, R.A.: The hallmarks of cancer. cell **100**(1), 57–70 (2000)
56. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. cell **144**(5), 646–674 (2011)
57. Krämer, A., Green, J., Pollard Jr, J., Tugendreich, S.: Causal analysis approaches in ingenuity pathway analysis. Bioinformatics **30**(4), 523–530 (2013)