Published in final edited form as:

IEEE Trans Biomed Eng. 2013 December ; 60(12): 3364–3367. doi:10.1109/TBME.2012.2212438.

omniBiomarker: A Web-Based Application for Knowledge-Driven Biomarker Identification

John H. Phan [Member, IEEE], Andrew N. Young, and May D. Wang* [Member, IEEE]

Abstract

We have developed omniBiomarker, a web-based application that uses knowledge from the NCI Cancer Gene Index to guide the selection of biologically relevant algorithms for identifying biomarkers. Biomarker identification from high-throughput genomic expression data is difficult because of data properties (i.e., small-sample size compared to large-feature size) as well as the large number of available feature selection algorithms. Thus, it is unclear which algorithm should be used for a particular dataset. These factors lead to instability in biomarker identification and affect the reproducibility of results. We introduce a method for computing the biological relevance of feature selection algorithms using an externally validated knowledge base of manually curated cancer biomarkers. Results suggest that knowledge-driven biomarker identification can improve microarray-based clinical prediction performance. omniBiomarker can be accessed at <http://omnibiomarker.bme.gatech.edu/>.

Section I Introduction

OmniBiomarker is a web-based application that uses a curated knowledge base of gene-to-cancer relationships for knowledge-driven analysis of high-throughput genomic data. Analysis of high-throughput genomic data has become a cornerstone of research aimed at understanding and treating diseases such as cancer, including renal [1], prostate [2], liver [3], and pancreatic [4] cancer. Molecular-based screening of cancer at an early stage can potentially improve diagnostic accuracy and therapeutic success. However, obstacles remain, both technical and practical, that prevent wide adoption of high-throughput molecular profiling in a clinical setting. Some of these obstacles include small-sample size, large-feature size, and numerous feature selection algorithms that can produce highly variable results [5], [6], [7]. Although many of these algorithms have successfully identified useful biomarkers, it is difficult to know, a priori, which algorithms will maximize the success of biomarker identification. We present an application that uses biological knowledge to guide algorithm selection in order to improve the robustness of biomarker identification.

We have previously shown that prior knowledge—clinically or biologically validated biomarkers—can improve the efficiency of biomarker discovery [8]. Knowledge integration can also alleviate the “curse-of-dimensionality” problem in high-throughput data analysis [9]. Several biomarker knowledge bases are available that can be used as a “ground truth”

May D. Wang is with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332. gtg538v@mail.gatech.edu, maywang@bme.gatech.edu.

for assessing the biological relevance of feature selection algorithms. Some of these include the GeneSigDB [10], the Gene Expression Atlas [11], the Cancer Gene Census [12], and the Cancer Gene Index (CGI) [13], [14]. We focus on the CGI as an initial knowledge base in omniBiomarker because it provides extensive links between genes and specific cancer subtypes in the NCI Thesaurus.

OmniBiomarker uses the CGI knowledge base to assign a biological relevance score to each feature selection algorithm. This score represents the probability that previously validated genes are favorably ranked compared to other genes. Researchers can consider this information to guide their choice of feature selection algorithm and, consequently, their choice of candidate biomarkers. Although various feature selection algorithms are widely used for prioritizing and identifying candidate biomarkers, these algorithms can be limited in their ability to identify important genes with low-differential expression and pronounced regulatory effects. For example, some genes may be involved in regulating the transcription of other genes. The protein products of these genes may be a part of signal amplification mechanisms such as dimerization or signaling cascades. Despite the inclusion of gene pathway and regulatory information in some knowledge bases, identification of novel interacting gene panels that can be used as disease biomarkers is still a major challenge in bioinformatics [7]. omniBiomarker currently integrates data and knowledge pertaining only to individual genes to assess the biological relevance of commonly used rank-based biomarker identification methods with the aim of increasing biomarker validation success [8].

In the following sections, we present an overview of the omniBiomarker system and a case study that examines the relationship between biological relevance of rank-based feature selection and microarray-based clinical prediction performance.

Section II System Design

omniBiomarker includes: 1) an extensive knowledge base of cancer subtypes and biomarkers; 2) algorithms that use this knowledge base to compute the biological relevance of feature selection methods; and 3) a web interface that simplifies knowledge-based gene expression analysis for both researchers and clinicians in search of cancer biomarkers. We describe these components of omniBiomarker in this section.

A. Biomarker Knowledge and Cancer Ontology

The omniBiomarker knowledge base contains 7213 cancer subtypes from the NCI Thesaurus and 6364 genes from the CGI [13], [14]. CGI genes are linked to the probes of 29 commonly used microarray platforms (Affymetrix, Agilent, and Illumina) by aligning the National Center for Biotechnology Information (NCBI) Reference Sequence of each gene to microarray probe sequences using the Burrows-Wheeler Aligner (BWA) algorithm [15]. Each CGI gene can be associated with several cancer subtypes defined in the NCI Thesaurus. Thus, the knowledge base contains extensive associations between cancer subtypes, genes, and microarray probes such that the interesting or important probes in a microarray dataset can be identified given the cancer subtypes of interest. The extent of links among the probes of 29 common microarray platforms and CGI genes associated with four

cancers can be observed at the omniBiomarker web site: <http://omnibiomarker.bme.gatech.edu/microarrays.php>.

B. Computing Biological Relevance

The biological relevance of a feature selection (i.e., gene ranking) algorithm is the probability that genes known to be important (i.e., differentially expressed) are ranked more favorably than genes that are unimportant to the disease of interest [8]. Given a gene expression dataset with N samples, we define a gene i as the vector

$\vec{d}_i = \left[(y_1, x_1^i), (y_2, x_2^i), \dots, (y_N, x_N^i) \right]$, where y_j is the class label of sample j and x_{ij} is the expression value of gene i of sample j . For each gene, we assign a rank using the function

$r_i = R_\theta(\vec{d}_i)$, where $\theta = \Theta$ is a meta-parameter representing the feature selection method. A smaller feature selection rank indicates a more differentially expressed gene. We define Θ as $\Theta \in \{\text{FC}, \text{T}, \text{SAM}, \text{RS}, \text{RP}, \text{mRMRD}, \text{mRMRQ}\}$, where FC = fold change, T = t-test, SAM = significance analysis of microarrays [16], RS = Wilcoxon rank sum test, RP = rank products [17], and mRMRD/Q = minimum redundancy maximum relevance with difference/quotient [18].

Given a gene expression dataset with M genes and a set of K knowledge genes $G_K = \{g_1, g_2, \dots, g_K\}$ where $g_k = 1 \dots M$, we define the biological relevance of a feature selection algorithm as

$$\phi(G_K, \theta) = \frac{1}{K(M-K)} \sum_{i \in G_K} \sum_{j \notin G_K} I(R_\theta(\vec{d}_i) < R_\theta(\vec{d}_j)) \quad (1)$$

Where $i = 1 \dots M, j = 1 \dots M, i \in G_K, j \notin G_K$, and $I(x)$ is the indicator function that evaluates to one when x is true and zero when x is false.

C. Web Interface

The omniBiomarker web application provides an interface for uploading gene expression datasets [see Fig. 1(a)] and associating these datasets to cancer subtypes. Users can interactively assign cancer subtypes in the NCI thesaurus to dataset labels by using an interactive cancer subtype browser [see Fig. 1(b)] or a keyword search tool [see Fig. 1(c)]. Fig. 1(d) is an example of four renal tumor subtypes assigned to dataset class labels. Once uploaded, the system automatically identifies knowledge genes from the CGI to be used for knowledge-driven analysis.

After uploading data and assigning data-specific diseases, users can select from seven common methods for ranking genes, as defined in Section II-B. Results of finished jobs can be viewed within the omniBiomarker web interface or downloaded. Downloadable results include ranked gene lists for each algorithm and each dataset. Additional information in the downloaded result includes, for each algorithm, the rank of each microarray probe, the score of the probe (which depends on the ranking algorithm), and the false discovery rate, when

applicable. Additionally, the biological relevance score for each algorithm is included in the table.

Section III Case Study

One of the goals of gene expression analysis is identification of biomarkers for diagnostic or prognostic medical applications. We examine the clinical utility of omniBiomarker by presenting a case study to determine the relationship between the biological relevance of feature selection and microarray-based clinical prediction performance. The case study analyzes several datasets across four endpoints: renal tumor subtype, prostate cancer, liver cancer, and pancreatic cancer.

A. Study Design

Predictive modeling with gene expression data involves optimization, or selection, of model parameters using training data (usually by performing cross validation) and evaluation of the model using independent testing data [19]. The process for determining model parameters typically includes choosing: 1) feature selection method; 2) feature size; and 3) classifier (see Fig. 2). We focus on determining if there is a relationship between the biological relevance of feature selection and predictive performance. Specifically, we want to know if prediction performance changes in the following scenarios: 1) feature selection using a method with high biological relevance versus that of a method with low biological relevance and 2) choosing a feature selection method based on biological relevance versus choosing one based on the training data alone (i.e., cross validation, data-driven). We examine the effect of feature selection on predictive performance under varying conditions of either optimizing the feature size and classifier, or averaging the performance of predictive models over all feature sizes and all classifiers (see Fig. 2).

As defined by (1), the biological relevance of a feature selection algorithm θ for a dataset l is $\phi_l(G_K, \theta)$ where $l = 1 \dots L$. Moreover, for each dataset, we can compute a cross validation score, $CV(\theta)$, and an external validation score $EV_{l,m}(\theta)$ on an independent dataset m , where $m = 1 \dots L$ and $l \neq m$. The cross validation score (5 folds, 10 iterations) for a dataset l and feature selection algorithm θ is the optimized (or averaged) prediction performance over feature sizes of 1 to 100 and over four classifiers: linear SVM (with cost ranging from 1 to 10), logistic regression, K nearest neighbors (with k ranging from 1 to 10), and linear discriminant (using pooled/unpooled and spherical/diagonal covariance). External validation scores are computed by optimizing or averaging prediction performance over a similar set of parameters (i.e., features sizes of 1 to 100 and four classifiers) but using dataset l for training and dataset m for testing. We use area under the receiver operating characteristic (ROC) curve (AUC) to measure prediction performance.

Given L datasets, we can compute the average prediction performance over all pairs of datasets in which one dataset is used for training and one dataset is used for testing. This average is

$$AUC = \frac{1}{L^2 - L} \sum_{\ell=1}^L \sum_{m \neq \ell} EV_{\ell,m}(\hat{\theta}) \quad (2)$$

where θ can be optimized using either a knowledge-driven or data-driven method. The knowledge-driven methods optimize θ using a biological relevance score:

$$AUC_{\max/\min} = \frac{1}{L^2 - L} \sum_{\ell=1}^L \sum_{m \neq \ell} EV_{\ell,m}(\arg \max_{\theta} / \min_{\theta} \phi_{\ell}(G_K, \theta)). \quad (3)$$

The data-driven method optimizes θ using cross-validation performance:

$$AUC_{\text{data}} = \frac{1}{L^2 - L} \sum_{\ell=1}^L \sum_{m \neq \ell} EV_{\ell,m} \left(\arg \max_{\theta} CV_{\ell}(\theta) \right). \quad (4)$$

We examine the distributions of $AUC_{\max} - AUC_{\min}$, $AUC_{\max} - AUC_{\text{data}}$, and $AUC_{\min} - AUC_{\text{data}}$ to assess the relationship between biological relevance of feature selection and prediction performance.

B. Biological Relevance of Feature Selection Affects Prediction Performance

Analysis of four renal (predicting tumor subtype), two prostate (cancer versus normal), three liver (cancer versus normal), and four pancreatic (cancer versus normal) cancer microarray datasets suggests that the biological relevance of feature selection methods affects prediction performance. These public datasets can be accessed from the omniBiomarker web site (http://omnibiomarker.bme.gatech.edu/data_public.php). Table I lists the AUC prediction performance for each cancer endpoint using different model selection methods. For the renal tumor, prostate cancer, and pancreatic cancer endpoints, there is an increasing trend in prediction performance from $AUC_{\min}$ (predictors optimized by choosing the least biologically relevant feature selection method, as defined by (3)) to AUC_{data} (predictors optimized using cross validation, as defined by (4)) to $AUC_{\max}$ (predictors optimized by choosing the most biologically relevant feature selection method, as defined by (3)). In Fig. 3 we examine the statistical significance of the trend by computing $AUC_{\min} - AUC_{\text{data}}$, $AUC_{\max} - AUC_{\text{data}}$, and $AUC_{\max} - AUC_{\min}$ for each of the four endpoints. We use a t -test to determine if the changes in prediction performance are significantly different from zero (i.e., p -value less than 0.05). The difference between $AUC_{\min}$ and AUC_{data} (see Fig. 3, orange boxplots) is statistically significant for the renal tumor and prostate cancer endpoints. The difference between $AUC_{\max}$ and AUC_{data} (see Fig. 3, green boxplots) is statistically significant for the renal tumor endpoint. $AUC_{\max}$ is significantly greater than $AUC_{\min}$ for the renal tumor and pancreatic cancer endpoints (see Fig. 3, blue boxplots). Although differences in prediction performance are not always statistically significant (e.g., liver cancer data), the trend in the data suggests a relationship between biological relevance of feature selection and prediction performance.

Section IV Conclusion

We have shown that the omniBiomarker knowledge base, consisting of genes from the CGI linked to cancer subtypes in the NCI Thesaurus, is comprehensive in terms of overlap with genes represented by microarray probes. Moreover, the genes in the knowledge base are distributed across thousands of specific cancer subtypes, making omniBiomarker applicable to a wide variety of cancer-related gene expression studies. We have also shown that the omniBiomarker knowledge base, when used to guide feature selection, can improve clinical prediction performance when applied to several microarray datasets across four-cancer types.

Quantitative evidence suggests a relationship between prediction performance and the biological relevance of feature selection. Case study results may change as we address some limitations of omniBiomarker. These limitations include 1) the subjectivity of gene-to-disease relationships in the knowledge base; 2) the limited set of feature selection algorithms tested; and 3) the lack of annotated biomarkers for predicting disease conditions such as treatment outcome. These limitations will likely be addressed as omniBiomarker evolves to refine the knowledge base and to expand the available set of feature selection algorithms. Future directions of omniBiomarker may include 1) similarity measures in the knowledge base to infer the functions of novel biomarkers or 2) a framework for measuring or improving the stability or reproducibility of feature selection algorithms with respect to biological knowledge.

References

1. Schuetz AN, Yin-Goen Q, Amin MB, Moreno CS, Cohen C, Hornsby CD, Yang WL, Petros JA, Issa MM, Pattaras JG, Ogan K, Marshall FF, Young AN. Molecular classification of renal tumors by gene expression profiling. *J Mol Diagn*. 2005; 7:206–218. [PubMed: 15858144]
2. Chandran U, Dhir R, Ma C, Michalopoulos G, Becich M, Gilbertson J. Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors. *BMC Cancer*. 2005; 5 art no. 45.
3. Roessler S, Long EL, Budhu A, Chen Y, Zhao X, Ji J, Walker R, Jia HL, Ye QH, Qin LX. Integrative genomic identification of genes on 8p associated with hepatocellular carcinoma progression and patient survival. *Gastroenterology*. 2011; 142:957–966. [PubMed: 22202459]
4. Badea L, Herlea V, Dima S, Dumitrascu T, Popescu I. Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology*. 2008; 55:2015–2026.
5. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Academy Sci USA*. 2006; 103:5923–5928.
6. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet*. 2005; 365:488–492. [PubMed: 15705458]
7. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007; 23:2507–17. [PubMed: 17720704]
8. Phan J, Yin-Goen Q, Young A, Wang M. Improving the efficiency of biomarker identification using biological knowledge. *Proc Pac Symp Biocomput*. 2009; 14:427–38.
9. Phan J, Moffitt R, Stokes T, Liu J, Young A, Nie S, Wang M. Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. *Trends Biotechnol*. 2009; 27:350–358. [PubMed: 19409634]
10. Culhane AC, Schröder MS, Sultana R, Picard SC, Martinelli EN, Kelly C, Haibe-Kains B, Kapushesky M, St Pierre AA, Flahive W. GeneSigDB: A manually curated database and resource

for analysis of gene expression signatures. *Nucleic Acids Res.* 2012; 40:D1060–D1066. [PubMed: 22110038]

11. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, Holloway E, Klebanov A, Kryvych N, Kurbatova N, Kurnosov P, Malone J, Melnichuk O, Petryszak R, Pultsin N, Rustici G, Tikhonov A, Travillian R, Williams E, Zorin A, Parkinson H, Brazma A. Gene Expression Atlas update—A value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 2011; 40(D1):D1077–81. [PubMed: 22064864]
12. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. *Nat Rev Cancer.* 2004; 4:177–183. [PubMed: 14993899]
13. Schueller, C.; Fritz, A.; Torres, S.; Wenger, K.; Albermann, K.; Komatsoulis, G.; Covitz, P.; Wright, L.; Hartel, F. Towards a comprehensive catalog of gene-disease and gene-drug relationships in cancer. Poster presented at the Thirteenth ISMB 2005 International Conference on Intelligent Systems for Molecular Biology; Detroit, MI. 2005; Available: <http://www.biomax.com/projects/nci.php>
14. Sioutos N, Coronado S, Haber M, Hartel F, Shaiu W, Wright L. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform.* 2007; 40:30–43. [PubMed: 16697710]
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2009; 25:1754–1760. [PubMed: 19451168]
16. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA.* 2001; 98:5116–5121. [PubMed: 11309499]
17. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol.* 2005; 3:1171–1190. [PubMed: 16278953]
18. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol.* 2005; 3:185–206. [PubMed: 15852500]
19. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L. The Micro Array Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol.* 2010; 28:827–38. [PubMed: 20676074]

Fig. 1. OmniBiomarker web interface for data upload and knowledge mining. (a) After uploading gene expression datasets, users can assign a cancer subtype to each class label. (b) omniBiomarker provides an interactive graphical search tool as well as (c) a keyword-based disease search tool for browsing the NCI thesaurus of cancer subtypes. (d) Assignment of specific cancer subtypes to class labels enables the system to automatically identify disease-specific genes for use in knowledge-driven biomarker identification.

Fig. 2.

Case study design. Predictive modeling with microarray data involves choosing (1) feature selection method, (2) feature size, and (3) classifier. Isolating the first step can elucidate the relationship between biological relevance of feature selection and prediction performance.

Fig. 3.

Biological relevance of feature selection affects microarray-based prediction performance. Prediction models with minimum biological relevance tend to perform worse than prediction models optimized with cross validation ($AUC_{min} - AUC_{data}$ tends to be less than zero, orange boxplots). Models with maximum biological relevance perform similarly to models optimized with cross validation ($AUC_{max} - AUC_{data}$ is close to zero, green boxplots) and perform better than models with minimum biological relevance ($AUC_{max} - AUC_{min}$ tends to be greater than zero, blue boxplots). P-values, computed using a t -test, indicate whether the changes in performance are significantly different from zero. Liver cancer results show no statistically significant trends.

TABLE I
AUC PREDICTION PERFORMANCE FOR FOUR CANCER ENDPOINTS USING DIFFERENT MODEL SELECTION METHODS

	Renal Tumor Subtype	Prostate Carcinoma	Liver Carcinoma	Pancreatic Carcinoma
AUC_{max}	0.96 (± 0.04)	0.76 (± 0.08)	0.91 (± 0.06)	0.72 (± 0.12)
AUC_{data}	0.94 (± 0.05)	0.73 (± 0.06)	0.93 (± 0.04)	0.71 (± 0.12)
AUC_{min}	0.91 (± 0.10)	0.69 (± 0.08)	0.93 (± 0.06)	0.67 (± 0.15)

* values indicate mean AUC ($\pm$ standard deviation).