

Pan-Cancer Analysis for Studying Cancer Stage using Protein Expression Data

Sameer Mishra, Chanchala D. Kaddi, and May D. Wang, *IEEE Senior Member*

Abstract— Pan-cancer analyses attempt to discover similar features among multiple cancers in order to identify fundamental patterns common to cancer development and progression. Pan-cancer analysis at the level of protein expression is particularly important because protein expression is more immediately related to patient phenotype than genomic or transcriptomic data. This study aims to analyze differentially expressed (DE) proteins between early and advanced cases of multiple cancer types through the usage of reverse-phase protein array data. The relevance of these proteins is further investigated by developing predictive models using K-nearest neighbor and linear discriminant analysis classifiers. The results of this study suggest that a pan-cancer analysis may be highly complementary to standard analysis of an individual cancer for identifying biologically relevant DE proteins, and can assist in developing effective predictive models for cancer progression.

I. INTRODUCTION

Cancer research has primarily been focused on studying the characteristics of a single cancer at a time. In contrast, the field of pan-cancer analyses aims to analyze the similarities and differences between multiple cancers simultaneously in order to better understand fundamental factors in cancer biology [1]. Previous pan-cancer studies have focused on genomic, epigenetic, transcriptional, and proteomic information, although protein expression information has been limited until recently [2]. The efforts of The Cancer Proteome Atlas (TCPA) have resulted in a comprehensive database of protein expression profiles across multiple cancers based off of reverse-phase protein array (RPPA) data [3]. However, there is still a gap in current knowledge on differentially expressed (DE) proteins and their roles in progression across multiple cancers.

Current research suggests that analyzing pan-cancer DE genes yields information on previously unconsidered genes relevant to the progress of individual cancers [4]. It may be

possible to find similar patterns at the protein level using pan-cancer DE proteins. This study aims to determine if DE proteins across multiple cancers have roles in cancer progression, and if these proteins can be used to build effective classification models to discriminate between cancer patients in early (stages I and II) and advanced (stages III and IV) stages of disease. The development of well-performing models would suggest that these DE proteins can be considered particularly relevant across multiple cancers, and may be investigated further to confirm that they are biomarkers.

II. METHODS

A. Data

Patient clinical information ($n = 3,202$) was obtained from The Cancer Genome Atlas (TCGA) for nine types of cancer. Corresponding protein expression data and a pan-cancer protein expression data set were obtained from TCPA. The pan-cancer data set contains normalized protein expression data for all patients in the individual cancer analyses. The nine types of cancer selected were bladder urothelial cancer (BLCA, $n=121$), breast cancer (BRCA, $n=906$), colon adenocarcinoma (COAD, $n=327$), head and neck squamous cell carcinoma (HNSC, $n=212$), kidney renal clear cell carcinoma (KIRC, $n=453$), lung adenocarcinoma (LUAD, $n=237$), lung squamous cell carcinoma (LUSC, $n=193$), ovarian cancer (OVCA, $n=407$), and uterine corpus endometroid carcinoma (UCEC, $n=346$). Patients were grouped based on their stage information, with Stage I and Stage II patients denoted as early stage and Stage III and Stage IV patients as advanced stage. The number of patients in each group for each cancer is shown in Table 1.

The proteins for which expression data is available were filtered by assessing the antibody validation status associated with each protein, as supplied by TCPA. 113 validated proteins out of the 187 original proteins are analyzed for both individual and pan-cancer analyses. This ensured that the data used for further analysis is from antibodies that have been determined as specific, selective, and reproducible [5].

B. Feature Selection and Classification Modeling

Figure 1 shows the overall workflow of this study. First, differential expression analysis was performed on both individual cancers and the combined pan-cancer dataset. The two-tailed t-test and Wilcoxon's rank-sum test were used to find DE proteins between the early stage and advanced stage patients, with a significance threshold of $\alpha = 0.05$. Bonferroni correction was then applied to control the false discovery rate. The relevance of DE proteins for each individual cancer was examined through peer-reviewed literature and the Human Protein Atlas [6]. These DE proteins were then compared to

This research has been supported by grants from The Parker H. Petit Institute for Bioengineering and Bioscience (IBB), Johnson & Johnson, Bio Imaging Mass Spectrometry Initiative at Georgia Tech, National Institutes of Health (Bioengineering Research Partnership R01CA108468, Center for Cancer Nanotechnology Excellence U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), Microsoft Research, the National Science Foundation (GRFP to CDK), P.E.O. International (Scholar Award to CDK), and the Georgia Institute of Technology Petit Undergraduate Research Scholars Program (Award to SM).

S. Mishra is with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: smishra39@gatech.edu)

C.D. Kaddi is with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: gtg538v@mail.gatech.edu)

M.D. Wang is with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (phone: 404-385-5059; e-mail: maywang@bme.gatech.edu).

the pan-cancer DE proteins in order to determine if pan-cancer DE analysis yielded any relevant proteins for a particular cancer type that the individual analyses had missed.

Second, in order to further assess the discriminatory relevance of identified proteins, classification models were developed using K-nearest neighbors (KNN) and linear discriminant analysis (LDA). Several alternative feature sets were used in these models: (i) all proteins; (ii) DE proteins identified through individual cancer analysis; (iii) DE proteins identified through pan-cancer analysis, (iv) proteins selected by mRMR (maximum relevance, minimum redundancy) on individual cancers, and (v) pan-cancer proteins selected by mRMR [7]. The mRMR method was implemented using the FEAST toolbox in MATLAB, and the five highest ranked features were implemented [8]. For the KNN models, alternative numbers of neighbors ($K = 1, 3, 5$, and 10) were tested.

The performance of each predictive model was evaluated

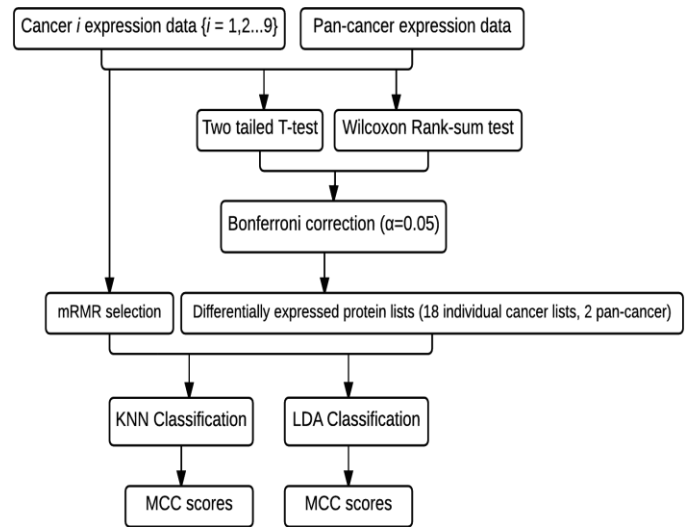


Figure 1: Workflow for performing pan-cancer analysis using protein expression data

III. RESULTS

Table 1: DE proteins and mRMR-selected proteins for each cancer type compared to DE proteins (Rank-sum test) from pan-cancer analysis. The “Selected unique” proteins represent proteins found only in the pan-cancer DE list, but which are known to be biologically relevant to the individual cancer type.

Cancer Type and Sample Information (Early, Advanced)	Individual Cancer			Pan-Cancer	
	T-Test	Rank-sum	mRMR	Common with Individual DE or mRMR	Selected Unique
Bladder urothelial cancer (BLCA) (36, 85)	None	None	Src_pY527, Raptor, MEK1, PEA-15, Rictor_pT1135	Raptor, MEK1, PEA-15, Rictor_pT1135	Bcl-2, N-Ras
Breast cancer (BRCA) (666, 240)	p_38_pT180_Y182	p_38_pT180_Y182	STAT5-alpha, PKC-delta_pS664, c-Jun_pS73, MIG-6, p27_pT198	p_38_pT180_Y182, PKC-delta_pS664, p27_pT198	Akt, mTOR, BCL2, ETS-1
Colon adenocarcinoma (COAD) (186, 141)	IRS1	None	MYH11, p38_MAPK, 4E-BP1, Shc_pY317, Smad4	MYH11, 4E-BP1, Smad4	ETS1
Head and neck squamous cell carcinoma (HNSC) (50, 162)	MAPK_pT202_Y204, MEK1_pS217_S221, S6_pS235_S236, S6_pS240_S244	MAPK_pT202_Y204, S6_pS235_S236, S6_pS240_S244	S6_pS240_S244, MEK1_pS217_S221, Dvl3, Cyclin_B1, Notch1	MAPK_pT202_Y204, MEK1_pS217_S221, S6_pS240_S244, Cyclin_B1	Akt, mTOR, N-Ras
Kidney renal clear cell carcinoma (KIRC) (263, 191)	4E-BP1, ACC_pS79, Akt_pS473, AMPK_pT172, Bad_pS112, beta-Catenin, CD31, Cyclin_B1, FASN, GAB2, HER2, HER3, MAPK_pT202_Y204, MEK1, MIG-6, p21, p70S6K_pT389, PEA-15, PI3K-p85, PTEN, Shc_pY317, Src, Src_pY527, TRFC, YB-1, YB-1_pS102	4E-BP1, ACC_pS79, Akt_pS473, AMPK_pT172, Bad_pS112, beta-Catenin, CD31, Cyclin_B1, ER-Alpha, FASN, GAB2, HER2, HER3, MAPK_pT202_Y204, MEK1, MIG-6, PEA-15, PI3K-p85, PKC-delta_pS664, PTEN, Shc_pY317, Src, Src_pY527, TRFC, YB-1, YB-1_pS102	Src_pY527, Bad_pS112, CD31, 4E-BP1, PI3K-p85	4E-BP1, AMPK_pT172, beta-Catenin, CD31, Cyclin_B1, MAPK_pT202_Y204, p21, PEA-15, PI3K-p85, PTEN,	Bcl-2, ETS-1
Lung adenocarcinoma (LUAD) (175, 62)	None	None	p38_pT180_Y182, PI3K-p85, EGFR, AR, PR	p38_pT180_Y182, EGFR, PR	14-3-3_zeta, 4E-BP1
Lung squamous cell carcinoma (LUSC) (157, 36)	None	None	MYH11, Shc_pY317, CDK1, PKC-delta_pS664	MYH11, PKC-delta_pS664	Bcl-2
Ovarian cancer (OVCA) (34, 373)	None	C-met_pY1235	MYH11, 14-3-3_beta, GATA3, Transglutaminase, Smad1	C-met_pY1235, MYH11, GATA3, Smad1	Akt, Beta-catenin
Uterine corpus endometroid carcinoma (UCEC) (268, 78)	E-Cadherin, ER-alpha	E-Cadherin, ER-alpha, PDK1_pS241	Transglutaminase, CD49b, PDK1, N-Cadherin, Tuberlin	ER-alpha, CD49b, PDK1	Cyclin B1

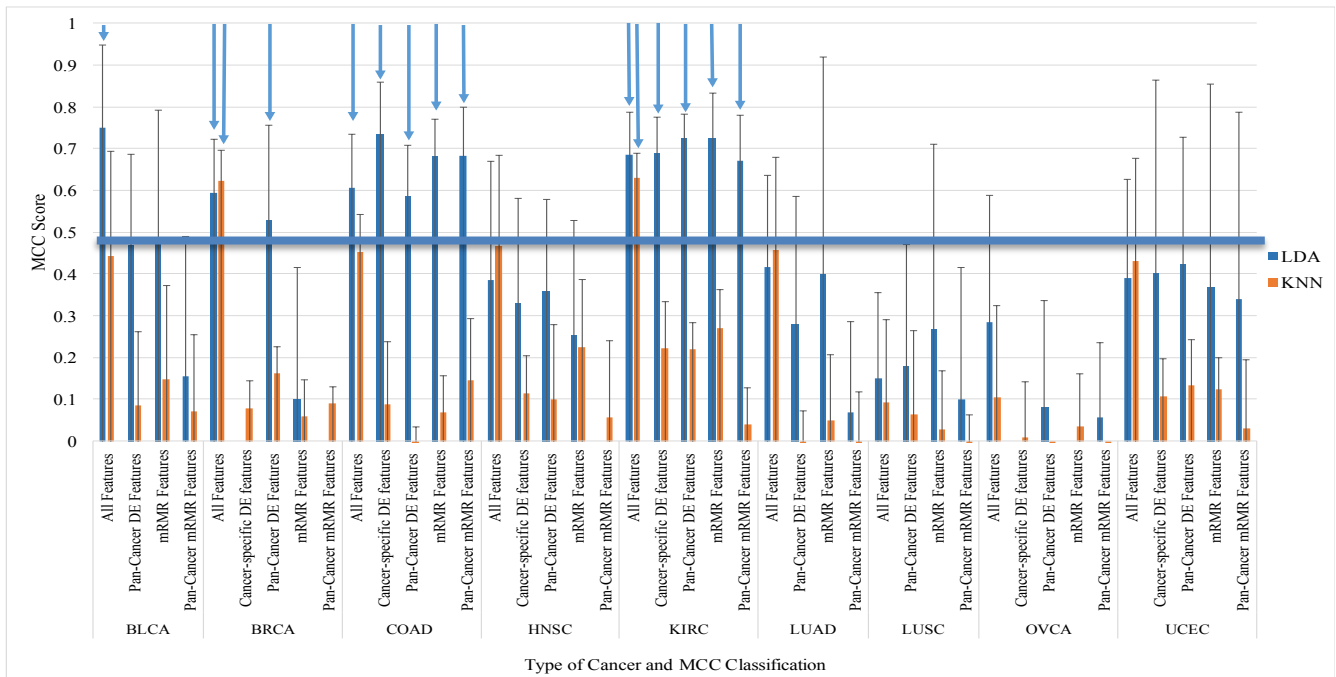


Figure 2: KNN and LDA Classification Model Performance

using Matthews Correlation Coefficient (MCC) as a metric. The mean MCC and standard deviation are reported following 10-fold cross-validation. MATLAB (MathWorks, Natick MA) was used for all analyses.

Table 1 shows the results of DE and mRMR analysis for each of the nine individual cancer types. For each individual cancer type, all significantly DE proteins are indicated, along with the top five selections from mRMR. In addition, these results are compared with the pan-cancer DE results. Proteins which are common between the pan-cancer DE protein list (not displayed in Table 1), and individual cancer DE lists and mRMR-selected features are shown. Also noted are selected proteins from the pan-cancer DE and mRMR feature lists that are relevant to each individual cancer, but do not appear in the individual cancer DE results.

Overall, pan-cancer analysis resulted in 56 DE proteins using the rank-sum test and 47 using the t-test. While the individual analysis of each cancer was able to find relevant DE proteins, the pan-cancer analysis was generally able to identify several of the same proteins, as well as additional DE proteins relevant to the progression of each cancer that were not identified in the individual cancer analyses. For example, the pan-cancer DE analysis identified several of the same proteins found by individual analysis for BRCA and HNSC, e.g., p_38_pT180_Y182 in BRCA and MAPK_pT202_Y204 in HNSC. In addition, pan-cancer analysis found Akt and mTOR, two proteins implicated in the progression of both BRCA and HNSC [9, 10]. Neither of these proteins was identified in the individual cancer analyses. In another example, no DE proteins were found for LUAD and LUSC. However, pan-cancer DE analysis identified several proteins, including p38_pT180_Y182, PR, and MYH11, found to be relevant to these cancers [11-15]. Literature also supports the roles of these proteins in different cancers, with p38_pT180_Y182 linked to BRCA and KIRC, PR linked to

BRCA, and MYH11 linked to BRCA and COAD [11, 15, 16]. These results suggest that pan-cancer DE analysis is helpful in identifying biologically relevant proteins that individual analysis may not because of limited sample size, lesser extent of differential expression, or other issues.

Next, the utility of the proteins identified through these alternative analyses was evaluated through predictive modeling. Fig. 2 shows the performance of each model across 10-fold cross-validation. For KNN, the results for $K = 1$ neighbor are shown, as poorer results were consistently observed for larger values of K tested for this dataset. Overall, several moderately performing models, with mean MCC values ≥ 0.5 , were developed and denoted with arrows in Fig. 2. The LDA models generally outperformed the KNN models, which may be due to the sensitivity of the KNN algorithm to noise [17]. Among alternative feature sets, the 'Pan-Cancer DE' and 'Cancer-specific DE' features appeared to give the highest, or close to highest performance for several cancers (COAD, KIRC, and UCEC). The 'Pan-Cancer mRMR' feature set generally resulted in worse performance than the 'Pan-Cancer DE', 'Cancer-specific mRMR', and 'Cancer-specific' feature sets, suggesting that mRMR is better suited to individual cancers for this type of data. However, for most cancer types, the best performance was observed with 'All Features', suggesting that incorporating more protein expression information improves performance.

Among the cancers for which DE proteins were identified through individual analysis (BRCA, COAD, HNSC, KIRC, OVCA, and UCEC), the 'Pan-Cancer DE' feature set yielded slightly improved performance over the 'Cancer-Specific DE' feature set for all except COAD. In BRCA, HNSC, and UCEC, this feature set also showed improved performance over the mRMR-selected features. These observations indicate that pan-cancer analysis identified relevant proteins that potentially have functional importance in many cancers.

IV. DISCUSSION AND CONCLUSION

This study investigates how pan-cancer analysis can be used to identify significant protein expression patterns across multiple cancers, and how predictive models can be developed using these pan-cancer results. Literature-based assessment of the results indicates that the pan-cancer DE approach successfully identified proteins which are biologically relevant in multiple cancers. Supervised analysis further demonstrated that the identified proteins were useful for developing discriminatory models for pathological stage.

Future research will expand upon these results by investigating other classifiers and feature selection methods, as well as determining an optimal number of features to select through mRMR and other filter-based feature selection methods. The statistical tests with Bonferroni correction were often unable to identify DE proteins for several cancers, which may be due to the conservative nature of the correctional method. An alternative correction method may yield additional relevant proteins.

The current results can be related to a recent pan-cancer study conducted on proteomic and TCGA-provided genomic information using unsupervised clustering analysis [18]. The clusters found in the study revealed proteins similarly expressed across multiple cancers. For example, Cluster V featured large sample populations across multiple cancers, including BRCA and KIRC, with elevated expression levels of MYH11, RICTOR, Caveolin1, and Collagen VI. Our current study compares well against the previous pan-cancer analysis by identifying all of the proteins highlighted in Cluster V as pan-cancer DE proteins. In addition, we apply supervised analysis with LDA and KNN classifiers to explore the performance of alternative feature sets in discriminating between early and advanced stage cancers. However, some differences remain for further investigation. For example, the previous pan-cancer analysis identified HER2 as biologically relevant across multiple cancers, while this study did not. Integrated analyses using multiple data types may yield additional insight into the importance of selected proteins across multiple cancer types, and a more comprehensive understanding of cancer biology.

In conclusion, pan-cancer analyses of protein expression data can lead to a better understanding of the fundamental factors in cancer biology, and can also assist in indicating potential therapeutic strategies and in patient stratification when validated by biologists and clinicians. Proteomic pan-cancer analysis has been limited due to the lack of publicly-available protein expression data. The TCPA initiative is a promising step in this direction. A larger set of proteins for analysis could help identify potential biomarkers and therapeutic targets. Pan-cancer proteomic analysis may also have implications towards therapeutics by suggesting cross application of drugs previously approved for one cancer type to other cancer types. Lastly, this study demonstrated that predictive models for cancer progression could be developed using pan-cancer protein expression data. Using pan-cancer protein expression data to develop models for other targets, including early diagnosis and patient survival, is an important direction for future research, and such models could serve as supportive tools for physicians.

ACKNOWLEDGMENT

The authors thank Dr. Chih-Wen Cheng, Po-Yen Wu, and Dr. John H. Phan for their assistance in preparing this manuscript.

REFERENCES

- [1] M. S. Cline, B. Craft, T. Swatoski, M. Goldman, S. Ma, D. Haussler, *et al.*, "Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser," *Sci. Rep.*, vol. 3, 10/02/online 2013.
- [2] N. The Cancer Genome Atlas Research, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat Genet*, vol. 45, pp. 1113-1120, 10/print 2013.
- [3] J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, *et al.*, "TCPA: a resource for cancer functional proteomics data," *Nat Meth*, vol. 10, pp. 1046-1047, 11/print 2013.
- [4] J. Reimand, O. Wagih, and G. D. Bader, "The mutational landscape of phosphorylation signaling in cancer," *Sci. Rep.*, vol. 3, 10/02/online 2013.
- [5] J. Bordeaux, A. W. Welsh, S. Agarwal, E. Killiam, M. T. Baquero, J. A. Hanna, *et al.*, "Antibody validation," *BioTechniques*, vol. 48, pp. 197-209, 2010.
- [6] F. Pontén, K. Jirstrom, and M. Uhlen, "The Human Protein Atlas—a tool for pathology," *The Journal of Pathology*, vol. 216, pp. 387-393, 2008.
- [7] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J Bioinform Comput Biol*, vol. 3, pp. 185-205, Apr 2005.
- [8] G. Brown, A. Pocock, M.-J. Zhao, M. Luj, and #225, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, pp. 27-66, 2012.
- [9] S. Bose, S. Chandran, J. M. Mirocha, and N. Bose, "The Akt pathway in human breast cancer: a tissue-array-based analysis," *Mod Pathol*, vol. 19, pp. 238-45, Feb 2006.
- [10] R. Vander Broek, S. Mohan, D. F. Eytan, Z. Chen, and C. Van Waes, "The PI3K/Akt/mTOR axis in head and neck cancer: functions, aberrations, cross-talk, and therapies," *Oral Diseases*, pp. n/a-n/a, 2013.
- [11] X. Wei, W. Guo, S. Wu, L. Wang, Y. Lu, B. Xu, *et al.*, "Inhibiting JNK Dephosphorylation and Induction of Apoptosis by Novel Anticancer Agent NSC-741909 in Cancer Cells," *The Journal of Biological Chemistry*, vol. 284, pp. 16948-16955, 2009.
- [12] Y. He, Z. Zhou, W. L. Hofstetter, Y. Zhou, W. Hu, C. Guo, *et al.*, "Aberrant Expression of Proteins Involved in Signal Transduction and DNA Repair Pathways in Lung Cancer and Their Association with Clinical Parameters," *PLoS ONE*, vol. 7, p. e31087, 2012.
- [13] H. Ishibashi, T. Suzuki, S. Suzuki, H. Niikawa, L. Lu, Y. Miki, *et al.*, "Progesterone receptor in non-small cell lung cancer—a potent prognostic factor and possible target for endocrine therapy," *Cancer Res*, vol. 65, pp. 6450-8, Jul 15 2005.
- [14] H. Jiang, Y. Deng, H.-S. Chen, L. Tao, Q. Sha, J. Chen, *et al.*, "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes," *BMC Bioinformatics*, vol. 5, p. 81, 2004.
- [15] E. Sebestyen, M. Zawisza, and E. Eyras, "Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer," *Nucleic Acids Res*, vol. 43, pp. 1345-56, Feb 18 2015.
- [16] C. K. Osborne, M. G. Yochmowitz, W. A. Knight, and W. L. McGuire, "The value of estrogen and progesterone receptors in the treatment of breast cancer," *Cancer*, vol. 46, pp. 2884-2888, 1980.
- [17] Y. Liu and G.-S. Chen, "KNN algorithm improving based on cloud model," in *Advanced Computer Control (ICACC), 2010 2nd International Conference on*, 2010, pp. 63-66.
- [18] R. Akbani, P. K. S. Ng, H. M. J. Werner, M. Shahmoradgoli, F. Zhang, Z. Ju, *et al.*, "A pan-cancer proteomic perspective on The Cancer Genome Atlas," *Nat Commun*, vol. 5, 05/29/online 2014.