Biomarkers from Biosimulations: Transcriptome-To-Reactome™ Technology for Individualized Medicine

Clyde F. Phelix, *Member, IEEE,* Greg Villareal, Richard G. LeBaron, and Dawnlee J. Roberson, *Senior Member, IEEE*

***Abstract*—We validated a model of the TGF-β signaling pathway using reactions from Reactome. Using a patent-pending technique, gene expression profiles from individual patients are used to determine model parameters. Gene expression profiles from 45 women, normal, or benign tumor and malignant breast cancer were used as training and validating sets for assessing clinical sensitivity and specificity. Biomarkers were identified from the biosimulation results using sensitivity analyses and derivative properties from the model. A membrane signaling marker had sensitivity of 80% and specificity of 60%; while a nuclear transcription factor marker had sensitivity of 80% and specificity of 90% to predict malignancy. Use of Fagan's nomogram increased probability from 7.5% for positive mammogram to 39% with positive results of the biosimulation for the nuclear marker. Our technology will allow researchers to identify and develop biomarkers and assist clinicians in diagnostic and treatment decision making.**

***Index Terms*— individualized medicine, diagnostic sensitivity, diagnostic specificity, simulation**

I. Introduction

THIS study was designed to follow the Phases of Discovery and Evaluation of Cancer Biomarkers [1], wherein the validated simulation model of the transforming growth factor (TGF) β-1 signaling pathway from Phelix et. al [2] was considered to have accomplished Phase I (pre-clinical exploratory studies) and Phase II (clinical assay/technique validation studies). This Transcriptome-To-Reactome™ Biosimulation Method uses gene expression levels to determine parameters in deterministic kinetic models of biochemical and signaling pathways [2]-[4]. The purpose of this brief report is to demonstrate the utility of this patent-pending Method as a tool that bridges the gap between biomarker discovery and clinical implementation. Because one patient's gene expression profile determines the parameters for one biosimulation, the Method will be used for Individualized Personalized Medicine. Unlike genomics that assign individuals to groups treated statistically [5], use of individual parameters from the patients for the mathematical modeling makes the predictive testing individualized [6]. Other forms of simulations integrating transcriptomics cannot achieve this standard yet [7].

As such, the gene expression profiles of peripheral blood mononuclear cells (PBMCs) from human females in categories of normal, benign, and malignant breast cancer were accessed from an existing public data archive [8], as accomplishing Phase III (retrospective validation studies for disease detection to evaluate sensitivity & specificity of disease detection). PBMCs from human cancer patients had been used to demonstrate the usefulness of an *ex vivo* **sti**mulation assay for assessing potential biomarkers of the TGFβ signaling pathway [9]. Human patients' PBMCs from GSE27562 [10] were the source for gene expression profiles in this study; and the TGFβ Signaling Model from Reactome [2] was used again to simulate an exposure to a bolus of TGFβ-1. Thus the Method is used as an '*ex vivo* ***sim****ulation* assay'.

This technology will advance discovery and development of biomarkers from benchtop to bedside and substantially reduce time to market for numerous biomarkers. This report is a first step in that translational, commercialization effort between academia and industry.

II. Procedure

A. Original Data Set

The sensitivity, specificity, and predictive values of the mammography test are known [11], and for the GSE27562 study [10]; where PBMC samples were collected from women with a suspect initial mammogram prior to undergoing a diagnostic biopsy procedure to determine whether the detected abnormality was benign or malignant. Blood was collected from women with a diagnosis of breast cancer, with a benign diagnosis, and with normal initial mammograms as negative controls. From their study, a total of 15 samples in each category were used for microarray gene expression profiles as our training data sets (n=5) and for the validation data sets (n=10). The PBMCs are also an interesting cell type because they are a

Manuscript received April 1, 2014. This work was supported in part by AL Phahelix Biometrics, Inc.

Clyde F. Phelix is with the University of Texas at San Antonio, TX 78249 and AL Phahelix Biometrics, Inc., San Antonio, TX 78269-1103 (corresponding author phone: 210-367-8953; fax: 210-458-5658; e-mail: clyde.phelix@utsa.edu)

Gregory Villareal is with AL Phahelix Biometrics, Inc., San Antonio, TX 78269-1103. (e-mail: gregvillarealphd@gmail.com).

Richard G. LeBaron is with the University of Texas at San Antonio, TX 78249 and AL Phahelix Biometrics, Inc., San Antonio, TX 78269-1103 (e-mail: richard.lebaron@utsa.edu)

Dawnlee J. Roberson is with AL Phahelix Biometrics, Inc., San Antonio, TX 78269-1103 USA (e-mail: dawn.roberson@alphahelixbiometrics.com).

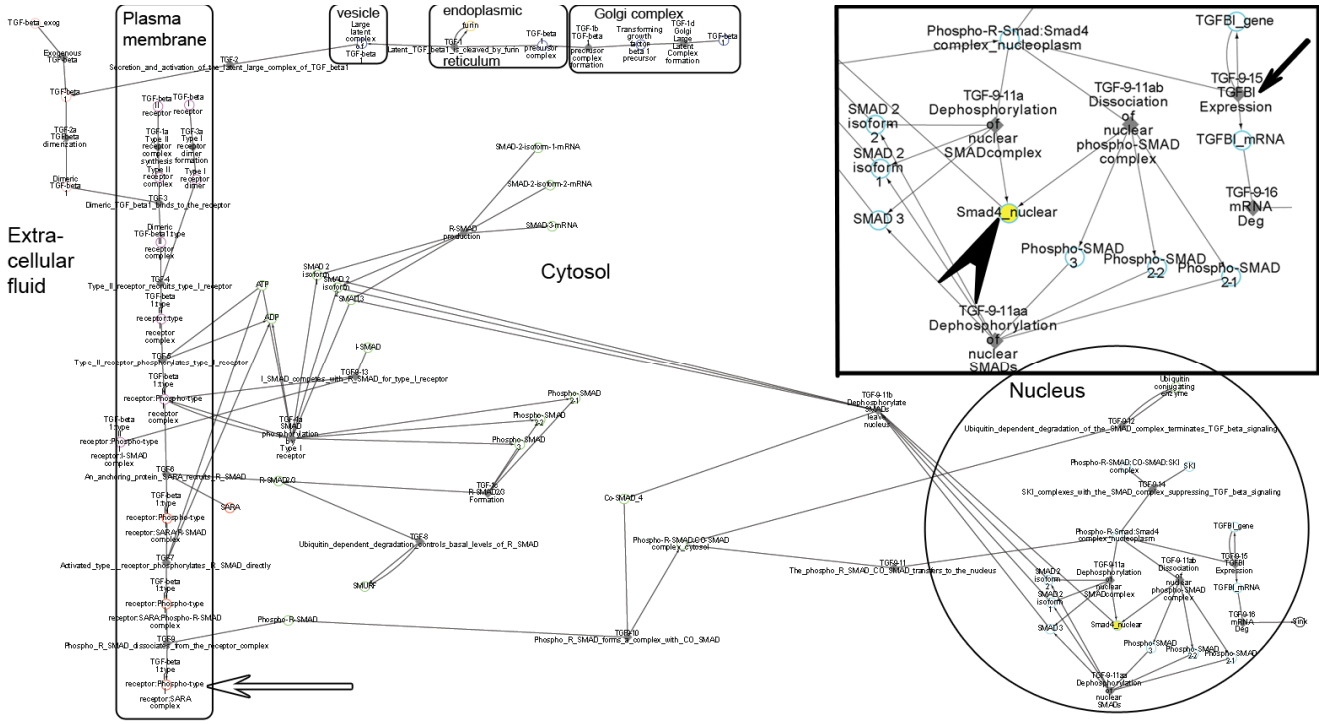

Fig. 1. Depicts the graphical display of the TGFβ-1 Signaling Pathway from Reactome. At top left the exogenous TGFβ-1 is administered to the extracellular space. The outlined arrow at bottom left indicates the 'TGF-beta-1-Type II receptor:Phospho-type I receptor:SARA complex' biomarker shown in Fig. 2c. The inset, enlarged from the nuclear compartment, shows the other two biomarkers considered in Figs. 3 (arrowhead) and 4 (arrow).

potential source of bone marrow mesenchymal stem cells that can infiltrate tumors and promote breast cancer metastasis [12] making the search for potential drug interventions of value [9].

Sensitivities analyses were used to identify biomarkers and candidate targets for novel drug development [13]. Also, because optimum biomarkers may also be a derivative property of the system, the slopes of temporal profiles for the reaction fluxes were assessed [14]. The generic 2 X 2 contingency table and formulae were used for calculating diagnostic sensitivity, specificity, positive predictive value, negative predictive value, prevalence, and likelihood ratios [15]. Standard methods for generating the receiver operating characteristic curve (ROC) and area under the ROC (AUC) were used [16]. A biomarker identified by sensitivities analysis was considered within only the groupings where the mammogram result was suspect, i.e., benign and malignant, as is often effective [17]-[19]. The second biomarker was evaluated by including all three categories.

Using the training set of PBMCs for assessing the "SARA" and Smad4 biomarkers identified by the sensitivities analyses in Fig.2, the mean values and standard deviation for peak response were calculated using sets of individual patient results in the three patient groups. Results of the validation data sets using the training data set results as cut off values for the "SARA" and Smad4 biomarkers test results were used to assign patients to the diagnostic categories of normal, benign, and malignant. The true positives, true negatives, false positives, and false negatives were calculated for patient test values. The 2 X 2 table [15] showed the calculations of sensitivity, specificity, positive predictive value, negative predictive value, and prevalence. Results of the training set of PBMCs for assessing the biomarker identified by the temporal analyses used the slope of the first 700 events that were calculated for each individual subject in the normal, benign, and malignant training data set. The results table showed the derivation of the cut off values for each bimarker and ranges used for the validation study, i.e., mean, plus or minus 3 standard deviations at increments of $1/35^{th}$. Results of the validation data sets using the training data set results as cut off values for the "slope of TGFBI mRNA expression flux" biomarker test results were used to assign patients to the diagnostic categories of normal, benign, and malignant. In this case, the calculation included normal, benign, and malignant patient cases, altogether and only as pairs of two.

B. Transcriptome-To-Reactome™ in silico model:

The TGF-β signaling pathway model was obtained from Reactome [20], downloaded as a SBML file that was imported into COPASI® [21]. Manual curation was required to adapt the model for the TGFBI_gene as a target for Smad transcription factor effector functions [22-26]. The curated model was imported into Cytoscape for imaging as a diagram (Fig.1) (http://www.cytoscape.org/).

The model incorporated 36 ordinary differential equations for 29 reactions (diamonds in Fig. 1) and 52

reactants or species (circles in Fig.1), using initial conditions determined from the genome-wide microarray data, GSE27562 [10] accessed from NCBI GEO [8]. There were 7 compartments (Fig.1). The model included 43 parameters derived by globalization [27] from expression levels of 35 genes for initial species levels and k-values of the mass action reactions. Genes and the derived k-values were assigned to reactions as in the Reactome SBML file or literature. The time course simulation as deterministic (LSODA) method was run to generate the species level and reaction flux value reports. The simulation duration was 6,000 model minutes. Additionally a sensitivities analysis was run on the time series simulation with the function as all variables of the model and variable as all parameter values. Microsoft® Excel® 2007 was used to assess and graph all data in the study.

III. Results

The biosimulations reanimated the PBMC response to 10 ng/ml of exogenous TGFβ-1 (Fig. 1), administered in a bolus dose, for each of the 45 individual patients. We generated a sensitivity analysis and performed a flux analysis for target gene expression, molecular analyses not previously available as molecular diagnostic tests.

A. Sensitivities Analyses of Biosimulations

Figure 2 shows bar graphs revealing the reactions that are sensitive to the various species in the model. With the naked eye, the normal and benign cases are nearly identical, but the malignant cases display heightened sensitivity to many of the species. Any of these species or reactions can be assessed for utility as a biomarker. Two were selected based upon the bars in Fig. 2c, one of which is indicated by the arrow.

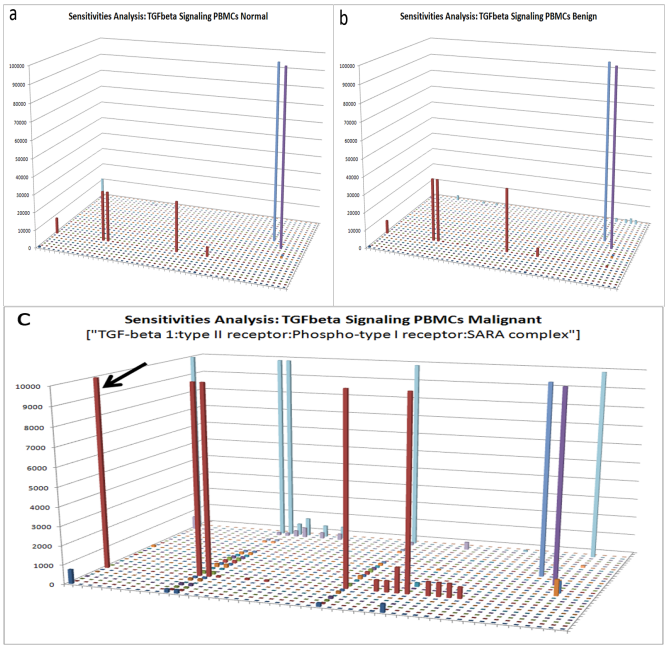

Fig. 2. Depicts the 3D graphical display of the sensitivities analyses results on the PBMCs from the normal (a), benign (b), and malignant (c) groups of patient subjects. These analyses represent the average for these groups from the training data set. Note the distinct appearance of the sensitive reactions (z-depth axis) to reactants (x- horizontal axis) in the malignant group. The arrow identifies a unique biomarker (TGF-beta-1-Type II receptor:Phospho-type I receptor:SARA complex). The reaction is the dissociation of Phospo-R-Smad from the activated Receptor complex. Because the biomarker is a complex of bound proteins with SARA being recruited to the activated receptors – and subsequently recruiting Smad-2 and Smad-3 to the receptor complex for phosphorylation, it is a candidate target for novel therapeutics.

B. Biomarker1 – species from the model

For the "SARA" biomarker, the sensitivity was 80%, specificity was 60%, positive predictive value was 67%, and negative predictive value was 75%, with a prevalence of 50%. Another species biomarker was tested, nuclear concentration of Smad-4 (Fig. 3), and at 85% accuracy the sensitivity was 80%, specificity was 90%, positive predictive value was 89%, negative predictive value was 82%, with a prevalence of 50%. The positive likelihood ratio was 8 and negative 0.22, and when using a known 7.5% probability of having breast cancer with a positive mammogram [10] on a Fagan's nomogram [28] the probability of having cancer is increased to 39% (Fig. 3b).

C. Biomarker2 – a derivative property of the system

Figure 4 shows the results of analyzing the slope of a flux curve for a target gene expression as a biomarker. The rate at which the flux of this gene expression event in the biosimulation reaches a maximum is obviously different for each of the classifications of patients in panel a. The assessments of the training and validation sets of patient biosimulations resulted in an AUC of 0.77 in panel b. At an accuracy of 85%, the sensitivity was 90%, specificity was 80%, positive predictive value was 82%, and negative predictive value was 89%, with a prevalence of 50%. The positive likelihood ratio was 4.5 and negative 0.12, and when using a known 7.5% probability of having breast cancer with a positive mammogram [10] on a Fagan's nomogram [28] the probability of having cancer is increased to 27%.

IV. Conclusions

A Method has been developed that uses microarray data as input to the model. This has been applied to detection and validation of biomarkers from a small publically available data set; a larger study would be more definitive.

We demonstrated utility of two approaches for

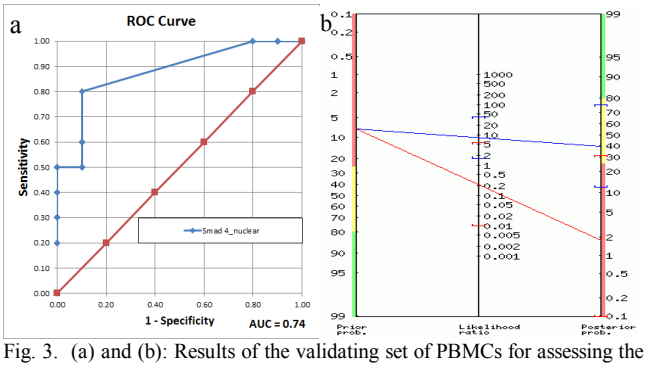

Fig. 3. (a) and (b): Results of the validating set of PBMCs for assessing the "Smad4" biomarker identified by the sensitivities analyses in Fig. 1. A ROC curve is shown in (a) where the area (AUC) is 0.74. (b) Fagan's nomogram shows clinical utility of the test [24].

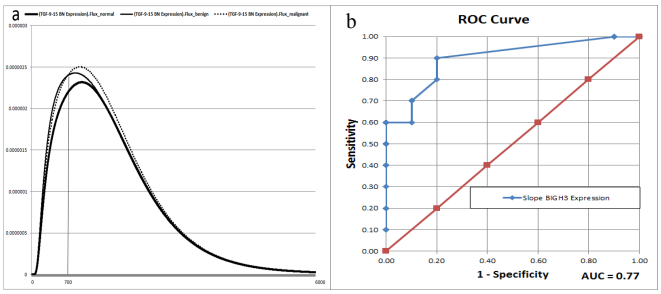

Fig. 4. (a) Temporal profile of the flux through the model simulation of the TGFBI (transforming growth factor beta induced protein ig-h3; also called BIGH3[26]) mRNA expression – the target gene of TGFβ1 signaling validated in [2]. These curves represent the averages of the training data sets for normal, benign, and malignant groups. The simulation time of 700 is shown with the vertical line that intersects with the first point of convergence of the benign (thin solid line) and malignant (dotted line) results. (b) Results of the validating set of PBMCs for assessing the biomarker identified by the temporal analyses. Slope of the first 700 was calculated for each individual subject in the normal, benign, and malignant training data set. The ROC curve shown is based upon differentiating malignant from benign.

identifying candidate biomarkers and then used routine methods for assessing their quality. A major advantage is that molecular assays and tests are not needed, saving time and money [1], [29], [30]. Overfitting of data is avoided unlike microarray results themselves [1], [13]. The approach is amenable to establishing multiple biomarkers that can enable screening to reduce false positive and negative cases [1]. The Method lends itself readily to drug development and targeted therapies, using all types of biomarkers [1].

V. Summary

Being able to view multiple species, their time course concentrations and their interactions provides insights into disease states that have not been available up until now with clinical test methods. This Methodology provides the opportunity to point to interactions of measures that can give medically actionable results for clinical utility. This technology can be scaled, is cost effective, and can be run on an individual, for precision medicine or to compare to average.

Acknowledgment

A UTSA patent is pending for the Method utilized in this study. AL Phahelix Biometrics, Inc. owns the trademark.

References

[1] Manne U, Srivastava, R-G, Srivastava S (2005) Recent advances in biomarkers for cancer diagnosis and treatment. Drug Discovery Today 1(14):965-976.

[2] Phelix, C.F., Watson, B., LeBaron, R.G., Villareal, G., Roberson, D. (2011) Transcriptome to Reactome deterministic modeling: Validation of *in silico* simulations of transforming growth factor-β1 signaling in MG63 osteosarcoma cells. CACS: Scientific Computing and Modeling 20(8):1-7.

[3] Phelix, C.F., LeBaron, R.G., Roberson, D.J., Villanueva R.E., Villareal, G., Rahimi, O.B., Siedlak, S., Zhu, X., Perry, G. (2011) *In vivo* and *in silico* evidence: Hippocampal cholesterol metabolism decreases with aging and increases with Alzheimer's disease. IEEE ICDM Biological Data Mining Applications in Health 21(9):1-7.

[4] Phelix, C.F., LeBaron, R.G., Roberson, D.J., Villanueva, R.E., Villareal, G., Rahimi, O.B., Siedlak, S., Zhu, X., Perry, G. (2011) Transcriptome-To-Metabolome™ Biosimulation reveals human hippocampal hypometabolism with age and Alzheimer's disease. Int. J. Knowledge Discovery in Bioinformatics 2(2):1-18.

[5] Penders B, Horstman K, Saris WHM, Vos R (2007) From individuals to groups: a review of the meaning of 'personalized' in nutrigenomics. Trends Food Sci & Technology 18:333-338

[6] Kronik N, Kogan Y, Elishmereni M, Halevi-Tobias K, et al. (2010) Predicting outcomes of prostate cancer immunotherapy by personalized mathematical models. PLoS ONE 5(12)e15482.

[7] Machado D and Herrgard M (2014) Systematic evaluation of methods for integrating transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol 10(4):e1003580.

[8] Barrett T et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37:D885-90.

[9] Baselga J et al. (2008) TGF-β signaling-related markers in cancer patients with bone metastasis. Biomarkers 12(2):217-236.

[10] LaBreche HG, Nevins JR, Huang E (2011) Integrating factor analysis and a transgenic mouse model to reveal a peripheral blood predictor of breast tumors. BMC Medical Genomics 4:61.

[11] Kavanagh AM et al. (2000) The sensitivity, specificity, and positive predictive value of screening mammography and symptomatic status. *J Med Screen* 2000 7:105-110.

[12] Karnoub AE et al. (2007) Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. Nature 449:557-565.

[13] van Riel NAW (2006) Dynamic modeling analysis of biochemical networks: mechanism-based models and model-based experiments. Briefings in Bioinformatics 7(4):364-374.

[14] Gomes B (2010) Practical applications of systems biology in the pharmaceutical industry. Int. Drug Discov. April/May 54-57.

[15] Spitalnic S. (2004) Test Properties I: sensitivity, specificity, and predictive values. Hospital Physician 27-31; www.turner-white.com.

[16] Spitalnic S. (2004) Test Properties II: Likelihood ratios, Bayes' formula, and receiver operating characteristic curves. Hospital Physician, October 2004, p. 53-58; www.turner-white.com.

[17] Mayeux R. (2004) Biomarkers: potential uses and limitations. The J Am Soc Exp NeuroTherapeutics 1:182-188.

[18] Maruvada P, Srivastava S (2004) Biomarkers for cancer diagnosis: Implications for nutritional research. J Nutr 134:1640S-1645S.

[19] Sharma S et al. (1999) Exclusion criteria enhance the specificity and positive predictive value of NMP22 and BTA STAT. J Urol 162:53-57.

[20] Croft D et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39:D691-7 Access 4/15/2010.

[21] Hoops S et al. (2006) COPASI—a COmplex PAthway SImulator. Bioinformatics 22(24):3067-3074.

[22] Zi Z, Klipp E. (2007) Constraint-based modeling and kinetic analysis of the Smad dependent TGF-β signaling pathway. PLoS One 2(9):e936.

[23] Clarke DC, Liu X. (2008) Decoding the quantitative nature of TGF-β/Smad signaling. Trends Cell Biol 18(9):430-442.

[24] Melke P, Jonsson H, Pardali E, ten Dijke P, Peterson C. (2006) A rate equation approach to elucidate the kinetics and robustness of the TGF-beta pathway. Biophysical Journal 91(12):4368-4380.

[25] Chung SW, Miles FL, Sikes RA, Cooper CR, Farach-Carson MC, Oqunnaike BA. (2009) Quantitative modeling and analysis of the transforming growth factor beta signaling pathway. Biophysical Journal 95(5):1733-1750.

[26] Yuan C, Yang M-C, Zins EJ, Boehlke CS, Huang AJ. (2004) Identification of the promoter region of the human βIGH3 gene. Molecular Vision 10:351-360.

[27] Fundel K, Kuffner R, Aigner T, Zimmer R. (2008) Normalization and gene p-value estimation: issues in microarray data processing. Bioinform Biol Insights 2:291-305.

[28] Caraguel CGB, Vanderstichel R (2013) The two-step Fagan's nomogram: ad hoc interpretation of a diagnostic test result without calculation. Evid Based Med 18(4):125-128; also see http://araw.mede.uic.edu/cgi-bin/testcalc.pl

[29] Taylor JMG, Ankerst DP, Andridge RR (2008) Validation of biomarker-based risk prediction models. Clin Cancer Res 14(19):5977-5983.

[30] Sturgeon C, Hill R, Hortin GL, Thompson D (2010) Taking a new biomarker into routine use – A perspective from the routine clinical biochemistry laboratory. Proteomics Clin Appl 4:892-903.