Diagnostic Proteomics: A New Approach

- **Focusing on the Tumor Microenvironment (TME) to Enhance Detection and Diagnosis**
- **Using Math and Physics to Suppress Proteomic Noise**

Keith Lingenfelter, Galina Krasik, Richard Saul PhD
OTraces Inc.
May 15, 2017

INTRODUCTION

This paper discusses a methodology which focuses on the Tumor Microenvironment as a vehicle for improved cancer detection and diagnosis and uses math- and physics-based noise suppression methods to improve the accuracy of human biological testing as measured by predictive power[1].

In recent years, cancer immunotherapy research has become increasingly interested in the Tumor Microenvironment (TME) which provides an ideal R&D platform for the development and evolution of new therapies and represents a potentially vast storehouse of diagnostic content[2]. The TME, which is bathed in the tumor interstitial fluid (TIF), is the cellular environment in which the tumor exists, including surrounding blood vessels, immune cells, fibroblasts, bone marrow-derived inflammatory cells, lymphocytes, signaling molecules and the extracellular matrix.

The TIF is also the transport fluid linking the tumor (and the TME) to the blood supply, and is important as it is the "battlefield messenger" for the active proteins that the immune system uses to try to suppress the tumor or the tumor expresses to assist its growth. These competing proteins, or cytokines, which are constantly at war with one another, fall into several functional categories of low level signaling proteins: pro- and anti-inflammatory, anti-tumor genesis (or cell apoptosis), angiogenesis and vascularization.

Although recognized as a potential source of rich diagnostic information, development of TIF analysis as a cancer screening modality has not progressed as sampling this fluid is very difficult and in order to do so means that the location of the tumor is known and therefore whether a tumor already exists. More challenging is detecting the presence of the TME/TIF and thus a malignancy without this knowledge. This requires a more accessible fluid for clinical diagnosis, such as blood serum, coupled with analysis of multiple proteins, known as proteomics, that may presumably be correlated to the presence or absence of disease. Serum presents some problem in this regard, as it is more an amalgam of the conditions in the patient's body than a direct pathway to detect the presence of an active TME (and thus a tumor).

In this white paper, we discuss a method for analyzing specific cytokines present in serum as an accurate proxy for the proteins active in the TME and TIF. The method involves several steps including two proprietary processes, termed proteomic noise suppression and multidimensional (or spatial) correlation. The method we describe can yield an accurate proxy for the actions of the proteins found in the TIF and thus is useful for detecting the presence of an active TME within the organism and thus a tumor. In essence, this method isolates the signature of the TME in the serum and indicates the presence (or not) of an active TME, indicating that an active tumor is present. Beyond this, the method measures the modulation of these proteins, which yields valuable information about the status of the tumor, degree of aggressive action and stage, as well as information about the immune system's progress in suppressing the tumor.

[1] Predictive power, the average of sensitivity and specificity, is a measure of a test's diagnostic reliability.

[2] The Tumor Microenvironment at a Turning Point— Knowledge Gained Over the Last Decade, and Challenges and Opportunities Ahead: A White Paper from the NCI TME Network; Yves A. DeClerck1,2, Kenneth J. Pienta3,4,5, Elisa C. Woodhouse6, Dinah S. Singer6, and Suresh Mohla6; *Cancer Res* 2017;77:1051-1059

ACCESSING THE TUMOR MICROENVIRONMENT

And the Burden Imposed by "Proteomic Noise"

In the particular case of cancer, a high degree of recent therapeutic research interest has focused on the so-called "tumor microenvironment" (TME) for development of treatment modalities[2]. Suppression or enhancement of the regulation of proteins active in the tumor interstitial fluid (TIF) found within the TME is thought to be a fruitful development path for these treatments. Proteins in the TIF that have been found to be good indicators are generally from five functional cytokine groups: pro- or anti-inflammatory, anti-tumor genesis (or cell apoptosis), angiogenesis and vascularization. Measurement of the activity of these proteins can provide insight into tumor activity and therapeutic impact. For example, treatment modalities that promote or suppress the protein activity can be monitored in the TIF to determine efficacy.

While appropriate for therapeutic applications, where the cancer is known to exist, sampling the TIF for diagnostic purposes has not been pursued. As the presence of TIF (and that of a TME) means, by definition, that the patient has an active tumor with a known location, its use as a diagnostic tool is moot. Beyond this, accessing these proteins for diagnosis when present in other bodily fluids, such as serum or urine, has not been considered because up until now the proteomic noise problem (discussed below) has rendered them unusable.

Herein we propose a method for producing an accurate proxy for the TME activity that uses active proteins in the TIF, using an easily accessible proxy fluid – in this case serum *(other fluids, such as urine are possible).* It should be noted that serum is an amalgam of the conditions of the overall organism (termed "proteomic noise") and is not specific to the tumor. ***The methodology we propose also eliminates proteomic noise, allowing an accurate assessment of the patient's condition***.

The method discussed herein involves; 1) selecting active TIF proteins that are indicative of conditions in the TME, 2) measuring these proteins in the serum proxy, 3) suppressing the proteomic noise to cleanly `identify cancer-related activity in the proteins, 4) then performing a correlation method that amplifies the actions of these proteins in a multidimensional matrix, and 5) scoring the protein activity to indicate the presence or absence of cancer, and if present, its development stage. This is done first to create a training set, representative of the population as a whole, that serves as a yardstick against which individual samples are then compared to determine their status – either diseased or disease-free.

THE PROTEOMICS NOISE PROBLEM

Unknowable Complexity

Diagnostic medicine has long held the promise that proteomics, the measurement of multiple proteins that can be correlated to the disease state, would yield break-through disease diagnostic methods for which research heretofore has not produced simple, viable blood tests. Cancer and Alzheimer's are just two such diseases.

A major problem has been identifying the actions of proteins (or other biomarkers) that are indicative of the disease state from samples whose protein actions may be influenced by a multitude of extraneous factors related to other conditions or drugs (prescribed or not, e.g., alcohol) that affect the protein measurements. Within a large population with known disease and disease-free states that could be used as the basis of a model to assess the correlation of the protein actions for diagnosis, there exists hundreds, if not thousands, of conditions or drugs that affect the up or down regulation of the proteins of choice. ***These non-disease factors that affect the expression of the protein and/or its up or down regulation and introduce the potential for diagnostic error are herein referred to as "Proteomic Noise"***.

Figure 1 below presents a scatter diagram of two typical proteins, *VEGF* and *IL6*, important biomarkers in determining cancer status, as measure in 400 patients that have been diagnosed (via mammogram and biopsy) as either having breast cancer (red) or being cancer-free (blue). Note the poor discrimination between the disease and not-disease data points.

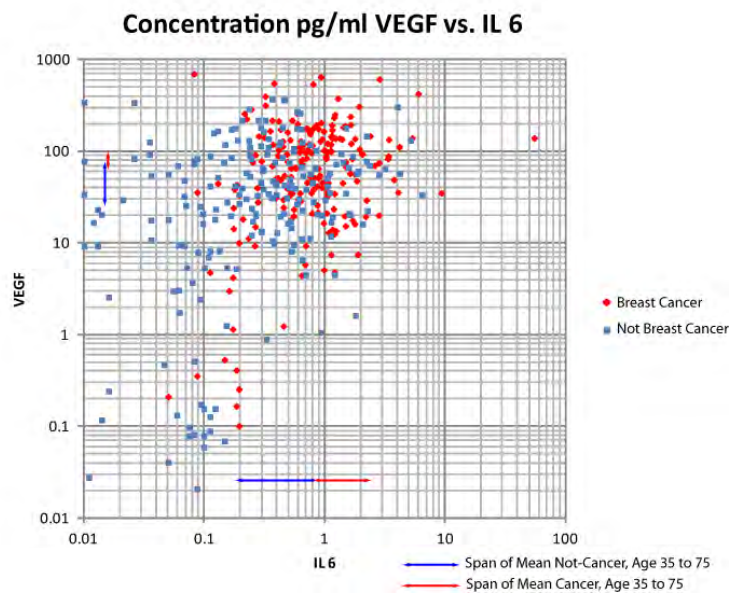

Figure 1. Two-protein scatter diagram

This plot is typical of hundreds of such plots with other biomarkers. ***This poor discrimination is endemic across almost all known biomarkers***.

There is some upward regulation of the biomarkers as the women transition from not-disease to disease, but the transition is clearly not well defined. This illustrates the problem of proteomic noise - most, if not all, of the samples most likely have conditions unrelated to breast cancer, some possibly known but mostly not. It is assumed that many are on drug therapies that may also affect the up or down regulation of these cytokines. Thus, the plot is "contaminated" with irrelevant information that confounds the correlation of these measured concentrations to the disease state[3].

In attempting to resolve this problem, proteomics research has tended to favor "big computation" methods to try to maximize the separation between the disease and not-disease states. These have tended to fall into four categories: 1. Regression methods, 2. Step-wise Receiver Operator Characteristic (ROC) curve area enhancements[4], 3. neural networks, and 4. what are called Support Vector Machines[5]. The well known regression and ROC curve methods simply are not capable of handling the data complexities illustrated in figure 1. Thus, research work has focused on the more complex computation methods:

- The **Neural Network** strategy inserts "neural" nodes between the inputs - biomarker concentration, and the outputs - the disease state. There are generally sufficient nodes such that each input has a unique pathway to each output through the neural network. Big computation then attempts to solve the correlation problem by assigning gain or attenuation (within the neural node) to each pathway for each input to each output.
- **Support Vector Machines** work by passing curved planes or surfaces through the complex biomarker plot space. This mathematical method is designed to find the optimal correlation separation surface between two states where the mixing of the training set data for the two states is high and this optimal surface is not discernible visually. The Support Vector Machine functions as a binary linear classifier that maps points in space with as large a separation (surface) as possible.

These two methods use what is termed a "training set" to arrive at a solution; data sets with known outcomes which attempt to impart intelligence into the complexity of proteomics. The theory is that if the training set-produced model is accurate, the model will be able to correctly classify unknown samples from the general population. While application of

---

[3] **The Complexity Paradox** (Kenneth L. Mossman, Oxford University Press, 2014), the challenges faced by Proteomic Investigators are aptly summarized: ***"the non-linear dynamics inherent in complex biological systems leads to irregular and unpredictable behaviors"***

[4] The Receiver Operator Characteristic (ROC) curve depicts the trade-off at the medical decision point (concentration level or cancer score) of true versus false negative and positive results. It is a visual representation of the tradeoff between these two important parameters.

[5] Computational intelligence techniques in bioinformatics; Aboul Ella Hassanien, Eiman Tamah Al-Shammari, Neveen I. Ghali; Computational Biology and Chemistry 47 (2013) 37– 47

the training set theory is used extensively in biologic applications, for reasons discussed below these methods have not been able to cut through the complexity typified in the Figure 1 plot.

The proteomic noise problem is further displayed in Table 1, which summarizes how various conditions or drugs affect the up or down regulation of the cytokines (and the tumor marker) used in a breast cancer detection panel[6]:

Table 1. Causes of Proteomic Variance (Noise)

Protein	Serum Concentration Regulation Direction	Condition or Drug	Incidence Approx.
IL 6	Up	Alcoholism	10.00%
		Asthma	10.00%
		Autoimmune disease	10.00%
		Bacterial Infections	0.50%
		Breast Cancer	0.50%
		Cardiomyopathy	0.01%
		COPD	10.00%
		Major depressive disorder (MDD)	-
		Pulmonary Hypertension	-
		Rheumatoid arthritis	2.00%
		Schizophrenia	3.00%
		Viral Infections (cold flu etc.)	8.00%
		Vaccines	1.00%
	Down	Certain Diets	-
		Corticosteroids Drug	-
		Flurbiprofen, Osteoarthritis Drug	-
		Immune suppressive disease	-
		Nonsteroidal anti-inflammatory drugs	-
		Tiaprofenic acid, Osteoarthritis Drug	-
		Valproic acid, epilepsy and bipolar disorder	-
IL 8	Up	Active inflammatory bowel disease	-
		Alcoholism	10.00%
		Breast Cancer	0.50%
		Chronic sarcoidosis	-
		Cushing's syndrome.	-
		Cystic fibrosis	-
		Drug Rexparixin	-
		Gingivitis	-
		Obesity	30.00%
		Psoriasis	-
		Pulmonary fibrosis	-
		Sepsis	-
		Thyroid diseases: goiter	-
		Viral Infections primarily Hepatitis	0.02%
		Vaccines	1.00%
	Up/Down	Menstrual Cycle	< 50 age
	Down	Corticosteroids Drug	-
		Immune suppressive disease	-
Kallikrein 3	Up	Breast Cancer	0.50%
		Alzheimer's disease	5.00%
		neural plasticity	-
		Psoriasis	-
		Skin desquamation	-
		Tooth development	-
TNFα	Up	Alcoholism	10.00%
		Alzheimer's Disease	0.50%
		Autoimmune disease	10.00%
		Breast Cancer	0.50%
		Graves Disease	-
		HIV	-
		Major depressive disorder (MDD)	-
		Psoriasis	-
		Severe heart failure	-
		Severe Septic Shock	-
		Some Parasites (malaria)	-
	Down	Antibiotics Some	-
		Flurbiprofen, Osteoarthritis	-
		Nonsteroidal anti-inflammatory drugs	-
		Tiaprofenic acid, Osteoarthritis	-
		Valproic acid, epilepsy and bipolar disorder	-
VEGF	Up	Age-related macular degeneration	-
		Asthma	10.00%
		Breast Cancer	0.50%
		Diabetes mellitus	10.00%
		Embryonic development	
		Heart disease	40.00%
		Muscle following exercise	High
		New blood vessels after injury	2.00%
		New vessels (collateral circulation) to bypass	-
		Rheumatoid arthritis	2.00%
		Simvastatin Drug	10.00%
		Stroke	1.50%
	Up/Down	Menstrual Cycle	< 50 age
		Anti-angiogenesis treatment for cancer Drug	-
		Dilative but not ischemic cardiomyopathy.	-
		Secondary progressive multiple sclerosis.	-
	Down	Thalidomide treatment for MS Drug	-

Legend - Markers Regulated per Condition or Drug	
	One marker per Condition
	Two markers per Condition
	Three markers per Condition
	All Five markers per Condition

Note that for just *VEGF* and *IL6*, there are thirty-five conditions listed. The legend below the table shows the combined number of markers for each condition. Yellow highlight indicates conditions or drugs that affect two of the proteins, tan indicates three, and light red indicates four or more affected proteins. **In this panel, only breast cancer affects four or more and in fact all five are affected, which further illustrates the proteomics complexity problem**.

[6] This table must be considered a very limited survey and, in fact, there are likely many unknown conditions or drugs (prescribed or not e.g., alcohol) that affect these protein concentrations in serum. Regardless, the noise suppression methodology discussed below will resolve these complexities.

PROTEOMIC NOISE SUPPRESSION AND SIGNAL RATIONALIZATION

The Solution from Measurement Mathematics

We have found that methods commonly used to dampen or suppress random or uncorrelated noise in physical science measurements can be applied in this situation to greatly improve the accuracy of the disease diagnosis.

The conventional wisdom in the disciplines of biology or clinical chemistry is that the "truth" lies in the measured raw concentration values. ***In contrast, the method described herein diverges completely away from that notion, and is based on a deeper interpretation of what the concentration values are telling us***. ***This approach dramatically improves on the performance of the traditional methods described above (regression, ROC curve enhancement, neural networks), renders the Support Vector Machine approach moot, and introduces a more powerful correlation method, termed "Spatial Proximity Correlation".*** The solution comes in part from the mathematics of measurement and the well-developed methodology used to reject random noise and produce the desired signal.

It is well-known in the physical sciences that all measurements consist of two elements, the desired signal and noise. By using mathematics***, it has been proven that the noise can be eliminated by multiple sampling of the desired signal to separate the noise into correlated (in sync with the measurement sampling scheme) and uncorrelated or random noise***. The random noise is then reduced by the square root of the number of samples. The signal and correlated noise (called offset) can be deduced very accurately by this multiple sampling technique. Finally, the offset can be determined with measurements in the absence of signal. As commonly used in communications technology, rapid multiple measurements of a single sample suppress noise by the square root of the number of samples, thus determining the actual signal to any degree of confidence desired.

The techniques described above are used in physics and communications; however, implementation in biology is decidedly different.

In biology (and proteomics), only a single sample of a patient (i.e., a single blood draw) is appropriate. However, by taking samples from many patients or subjects within a given population, the disease or disease-free signature of that population can be determined, and applying these characteristics to a single patient will allow determination of the disease state of that patient based upon this signature. The deeper understanding of concentration values starts with the notion that in and of itself the raw concentration value is not "truth" but is a combination of signal (disease), offset (not disease) and noise.

In the case of proteomics, the noise is fixed in time for any one sample (i.e., a single sample from an individual being tested for disease). The diagnosis must be made after one sample, not after months of sampling. Thus, a somewhat different strategy must be used, but the underlying mathematics is similar.

For proteomics, many hundreds of different sample measurements from individuals within known groups, disease and disease-free, are taken to determine the mean values of the signal (disease) and offset (disease-free). The accuracy of these parameters is only limited by the number of samples taken[7]. Once these mean values are determined, some rationality can begin to be applied to the Figure 1 scatter diagram.

This method cannot fully determine the accurate disease or disease-free values for a single sample from a particular individual, as the proteomic noise for any given sample is fixed at the time it is taken. However, a brief thought experiment illustrates that this parameter is only meaningful in the context of a large number of samples representative of the

[7] The number of samples taken is determined using conventional statistical sampling methodology.

population as a whole. For example, a particular individual must first have a disease to try to measure the mean value for that disease; therefore the disease-free mean value could not exist for this one individual. A baseline for just that individual could be established over a long period of time, but it would also be contaminated by the proteomic noise discussed above. Managing the impact of proteomic noise in one individual would be simplified; however, the mean value of any disease needs to be based upon a large population sample to be useful, again because of the proteomics noise problem. The useful information in this case are the mean values (either disease or disease-free) for the population in general, which can then be used to correctly diagnose that one individual. The methodology to achieve this is discussed below.

The first step is to reconcile what can be known about the Figure 1 scatter diagram. Our research has determined that there are only four useful pieces of information in the plot - the mean values of the two biomarkers for those samples which have cancer and those that are cancer free[8].

Using these mean values as a separator, we can then assign each sample to a quadrant. There are only four quadrants; each individual sample is either: 1) less than the mean value for all breast cancer free samples; 2) greater than this value but less than the derived mid-point mean value between the breast cancer/breast cancer-free means; 3) above this derived midpoint of the means but below the mean value for cancer; and 4) above the mean value for breast cancer. Any information beyond this for individual samples is not useful and can be considered noise[9]. Once assigned to a quadrant each individual sample can be rank ordered by its relationship to (i.e., distance from) the respective means.

There is a significant complication in this ranking and noise damping process - the mean values of the population vary dramatically with age. Thus, the mathematical method of placing these samples by quadrant, 1 through 4 above, must also sort out the age drift problem. This problem can be serious enough that the disease-free mean values will overlap the disease mean values at different ages in some cases. This drift must be zeroed out in the correlation. (The raw concentration values shown in figure 1 have embedded mean value age drift while in figure 2 below this drift is normalized).

Once normalized, a new independent variable based upon the age-normalized ranking and the damping of noise is called for. We term this new variable the "Proximity Score." The Proximity Score encompasses the above-noted attributes including: 1) being anchored by the disease and disease-free mean values; 2) age drift is normalized (zeroed-out) at the mid-point transition from disease-free to diseased state; 3) the individual samples are force ranked by their relationship to the means; and 4) the outlier “noise” in samples furthest from the means is mathematically dampened or compressed. In addition, the clustering behavior of the raw concentrations in the ‘noisier” outlier samples is retained and applied to the correlation methods, retaining their spatial relation, as discussed below.

Figure 2 below shows the Proximity Score plot for the same two biomarkers for the 400 women shown in Figure 1. In this plot, the age drift is normalized, and the mean values of the cancerous and breast cancer-free samples are now fixed at proximity scores of 4 and 16 for each biomarker respectively. (Proximity scores are arbitrary; for this example they were chosen to range from 0 to 20). The individual sample data points are forced into the ranking zones (1 through 4) inside the fixed mean values. At a fixed Proximity Score of 11, both biomarkers are at their derived mean point between the

[8] This is example is based on a breast cancer sample. The approach described is applicable to other disease types as well.

[9] Some physical scientists may object to the use of the term “noise” as noise is usually considered random. The proteomics noise discussed here is caused by generally unknowable actions or conditions (drugs, environmental factors or individual factors (e.g., genetic variations, etc.)). The “noise” can also be termed “Proteomics Variance.” However, since the conditions that cause these variances are so numerous and randomly distributed in the population, they can rightly be considered uncorrelated or random noise, and treated as such. This means that information contained in very far outlier concentrations measured in some samples, for example, is useless information and can be damped (suppressed mathematically).

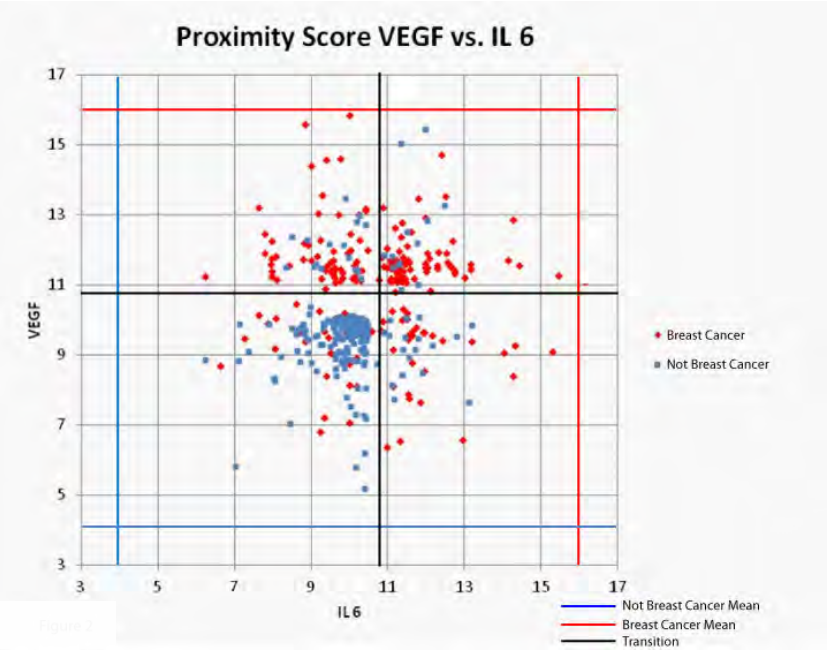

Figure 2. Two-protein age-normalized Proximity Score scatter diagram

breast cancer and cancer-free means. As age drift has been normalized, a raw sample exactly at the concentration of either of the means or the mid-point between the means will have a Proximity Score of 6, 11, or 16 respectively, regardless of age. All other individual samples including far outliers are compressed into the space between the means, and each raw concentration value is forced to the proper side of the mid-point of the means by the relationship of the raw concentration value to the means and mid-point of the means.

CORRELATION IMPROVEMENTS

Steps to Raising Predictive Power

When these new independent variables are applied to various correlation methods, the results are considerably improved. Table 2 shows the improvements in predictive power (the average of sensitivity and specificity), and more improvements are discussed below.

Table 2 - Predictive Power Improvement
by Conversion from Raw Concentration Scores to Proximity Score

Data Manipulation Method	Correlation Method	Predictive Power Score	Percentage Point improvement
Logarithm of raw concentration	Logistic Regression	80%	Baseline
	Neural Network	84%	4%
	Surface Vector Machine	84%	4%
Conversion of Concentration to Proximity Score	Logistic Regression	85%	5%
	Neural Network	87%	7%
	Surface Vector Machine	90%	10%
	Spatial Proximity	90%	10%
Conversion of Concentration to Proximity Score plus Orthogonal Biomarkers	Spatial Proximity	96%	16%

As can be seen from Table 2, simply converting to the Proximity Score from raw concentration improves regression methods by 5% and neural networks by 7%. Support Vector Machine results are improved by 10%. Using the OTraces-developed correlation method, Spatial Proximity Correlation (discussed below), alone will yield a similar improvement as the Support Vector Machine method.

Concentration Trends versus Clusters of Biomarkers

Regression methods and neural networks focus on data trends and cannot retain any spatial separation information. The evidence shows that predictive power improvements are enhanced by focusing on up/down regulation clustering in the multi-dimensional biomarker space rather than following data trending in concentration measurements, especially after the conversion from raw concentration data to Proximity Score. The Support Vector Machine and Spatial Proximity method captures this spatial separation information, discussed more below.

As noted earlier, the Support Vector Machine is a mathematical method designed to find the optimal correlation separation surface between two states where the mixing of the training set data for the two states is high and this optimal surface is not discernible visually. It functions as a binary linear classifier that maps points in space with as large a separation (surface) as possible. The OTraces-developed computation methods described herein will produce this separation by damping the "noise" and placing the planes of best separation on the multi-dimensional plot that can be seen with the eye (such as the midpoint at Proximity Score of 11 in Figure 2). Effectively this renders the Support Vector Machine results moot.

Proteins with "Orthogonal Functionality"

In the breast cancer example shown, the selected biomarkers have functions that are specific to the body's reaction to the disease or the disease's action on the body. In the case of cancer, these are generally considered to be active proteins such as inflammatory (pro or anti), cell apoptosis or vascularization functions. Many cytokines have multiple interacting functions. Thus, the task is to select functions and the proteins such that this interaction is limited or act independently of other biomarker functions. In other words, the varying levels of activity observed in any one biomarker should not interact or influence the activity of the others, **except as the disease itself affects the others.** Thus, if variations occur in one function, these changes in and of themselves should not drive changes in the others. We call these independent functions "orthogonal", as an orthogonal plot will separate out independent actions, improving predictive power. Including proteins with very similar functions will not demonstrate this orthogonal separation and thus will not improve correlation or result in predictive power gains, making the selection of appropriate cytokines critical.

Capturing Orthogonal Functionality in the Correlation

This functional orthogonal action of proteins (or other biomarkers) can easily be seen when they are plotted on orthogonal axes once proteomic noise is suppressed. If the biomarker up-regulates in the transition to a disease state, the positions of the biomarkers in the dimensional grid will move away from the ordinate, and will be obvious to the eye. This dimensional movement is dramatically enhanced by the conversion to Proximity Score. (in fact, when using other analytical techniques, the contamination by proteomics noise almost completely obscures this information, and is lost when regression or neural network correlation methods are used).

Capturing this information in a multi-dimensional grid is intrinsic to the correlation method. Both the Support Vector Machine and the Spatial Proximity methods do this. Referring to Figure 2 again, the surface of maximum separation for best correlation is at about Proximity Score 11, the derived midpoint of the means, for both biomarkers. If Support Vector Machine analysis were run on this Proximity Score plot, it would confirm the visual recognition of this plane. Thus using functional orthogonality coupled with the Spatial Proximity Correlation method on these complex functional cytokines yields significant improvement in predictive power. Note also that the Support Vector Machine does not specify how the actual correlation weighting is done, just the planes of maximum separation in the multi-dimensional plot. Spatial Proximity focuses first on clustering of the data then on data trending in the transition from disease-free to disease state.

The Spatial Proximity Correlation Method

This correlation method has not heretofore been used in proteomics likely because the software implementation is complex and unavailable as prepackaged "out of the box" software. The method consists of constructing a multi-dimensional computerized grid, one for each biomarker, using training set results as a scoring mechanism. The Proximity Score for each biomarker in the training set (which mimics the general population, based on known cancerous or cancer-free cytokine activity) is plotted on this grid (typically five dimensions (or axes) are utilized). Each grid space (whether it includes a training set data point or not) in this five-dimensional grid[10] is scored (as cancer or cancer-free) by its proximity to several (15 to 20) training set points plotted on the grid. The score is arbitrary (generally 0-200) based on its relative proximity to the training set data point, with a score of 200 being cancerous and 0 being cancer free. Biomarker results from patient samples (unknown as to cancer state) are evaluated by placing them on this "evaluation grid" and scored according to the score in each grid square into which the patient sample falls. As shown in the last line of Table 2, combining functional orthogonal selection of biomarkers with the Proximity Score conversion methodology (noise reduction and age normalization) yields predictive power improvement to 96%[11].

Figure 3 below presents the data in figure 2 on a three-dimensional plot where the vertical axis is the population distribution of each biomarker. The Proximity Score separates the sample data into two groups, populated by mostly cancer-free subjects (blue – close to the origin) and those with breast cancer (red - further away from the origin). Notice the single biomarker overlap displayed on each of the horizontal axes (*IL6* and *VEGF*). No amount of mathematical manipulation can eliminate this problem. Notice however, that individual red (Breast Cancer) samples that are low on the pro-inflammatory axis (*IL6*) tend to have a high position on the vascularization (*VEGF*) axis. The same behavior is true of the other horizontal axis (*VEGF* vs. *IL6*). This separation will occur when functionally orthogonal biomarkers are used, or with tumor markers that do not have inherent orthogonal separation characteristics. Simple probability will dictate that a low-level concentration for one of the tumor markers will very likely have high levels for all the others in a cancer patient. This separation effect is brought out when the proteomic variance or noise is dampened, as opposed to what is observed in the raw concentration values, where this separation effect is contaminated by noise.

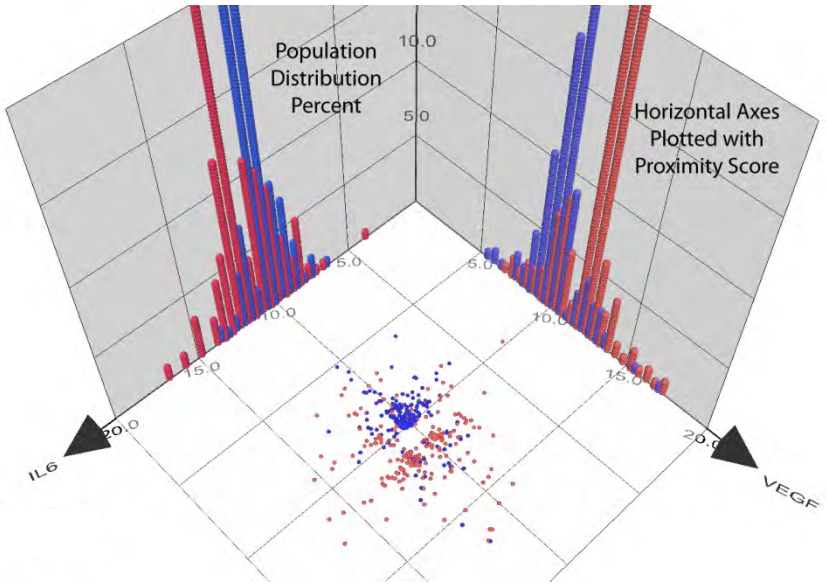

Figure 3. 3-D Proximity Score Plot

While this illustration shows only two biomarkers, this separation keeps accumulating through all five orthogonal dimensions in the grid, whether the biomarkers are chosen for orthogonality of function or are just tumor markers that potentially indicate the presence of a particular tumor.

[10] A typical five-dimensional grid model consists of $2{,}000^5$ or 32 quadrillion grid spaces.

[11] Results from a 400 person OTraces breast cancer trial.

Figure 4 below shows a third dimension (in this case adding *IL8*). In Figure 5, the 3-D plot is rotated around the vertical axis and up-tilted slightly to show the increased separation as axes are added. Adding the 4th and 5th orthogonal dimensions continues to refine this separation and amplifies the predictive power. An OTraces-developed proprietary computer algorithm processes through all 5 dimensions, ultimately yielding the amplified area under the Receiver Operator Characteristic (ROC) curves for breast cancer and prostate cancer discussed in the next section.

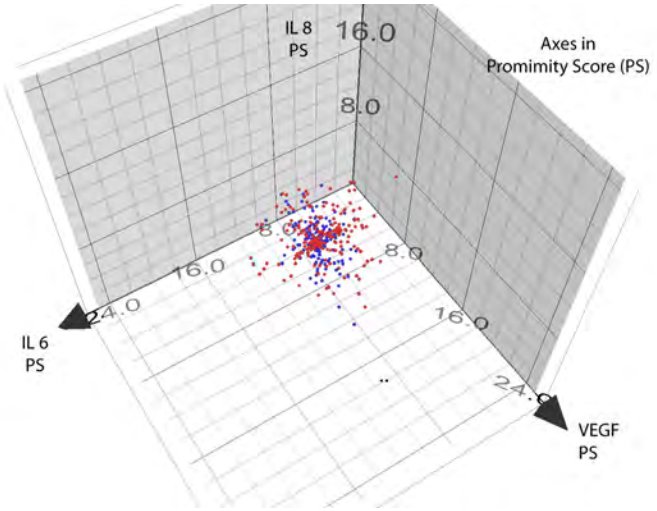

Figure 4. 3-D plot of 3 proteins – IL6, IL8 and VEGF

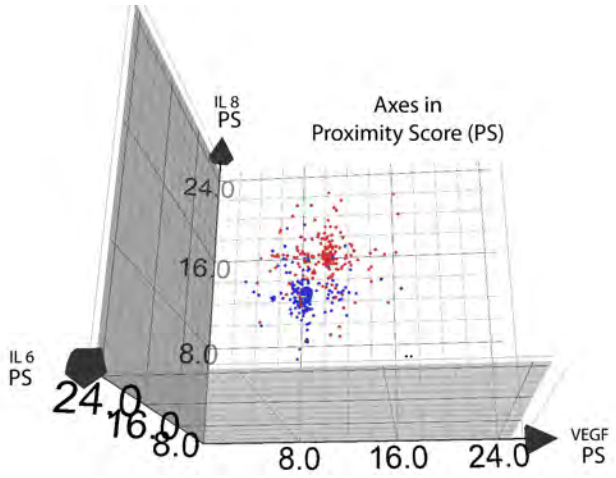

Figure 5. 3-D plot rotated about vertical axis and up-tilted

ROC CURVES – OTRACES TRIALS

The ROC curve is a common way to graphically compare the Predictive Power of a given test, wherein the true positive rate (Sensitivity) is plotted as a function of the false positive rate (1-Specificity) for different cut-off points of a medically relevant parameter.

Breast Cancer Results

The Breast Cancer Test Panel

The two biomarkers referred to above, *VEGF* and *IL6*, are part of a breast cancer diagnostic screening test produced by OTraces called *BC Sera Dx*. The full test panel is derived from the protein compliment active within the TIF from the TME. It consists of *IL6* (pro-inflammatory), *TNFα* (tumor necrosis factor), *IL8* (angiogenesis), *VEGF* (vascularization) and a tumor marker, *Kallikrein 3*[12] (This protein has no apparent function in the growth or demise of the tumor). The test panel produces predictive power of greater than 95% using the proteomic noise suppression methods described herein. The ROC curve for this test panel using these methods is shown in Figure 6, below. This high predictive power is the result of using of proteins derived from the TIF.

Note that these biomarkers will only produce about 80% predictive power when using the logistic regression analysis or the simplistic ROC curve area maximization method due to the proteomic noise problem described earlier.

[12] Kallikrein 3 is also known as Prostate Specific Antigen (PSA). It is used in the current FDA-approved prostate cancer screening test. It is also a tumor marker for breast cancer.

Figure 6 shows the resulting ROC curve for the same 400 known samples as in Figure 2, 50% with and 50% without breast cancer. The blood sample measurements were run on the OTraces CDx Chemistry System, at the Gertsen Institute in Moscow, Russia and analyzed with the proprietary OTraces Spatial Proximity algorithm which resides on a U.S. based server. ***The ROC curves show the buildup of predictive power as biomarkers are added. The highest curve (shown in red) shows the combination of the five dimensions of biomarkers achieved using the OTraces approach. This curve achieves over 98% AUC.*** The amplification results directly from the suppression of proteomic noise and the amplification effect from the retention of the "spatial" separation from the spatial proximity correlation method.

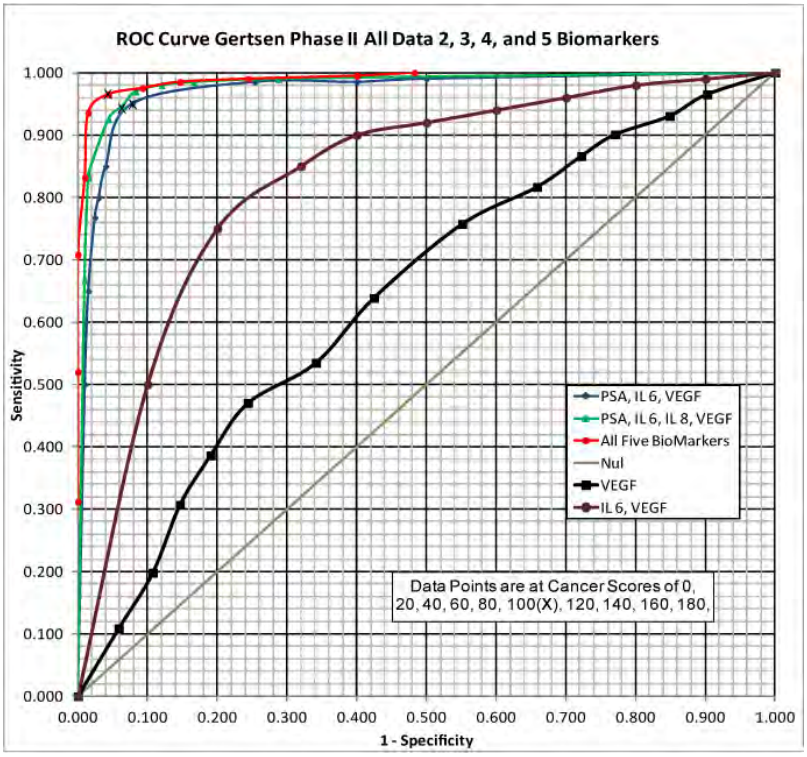

Figure 6. ROC Curve for 400 breast cancer samples.
Source: The Gertsen Institute test data

The test cohorts in this project were as follows:

Cancer-free Cohort **-** 200 samples drawn from anonymous sources. They were not prescreened for being cancer-free, so there is a small possibility of any one sample actually having breast cancer. The makeup of the cohort was designed to mimic a general population of women who would present for an annual screening mammogram and abnormal conditions not related to cancer but present in the general population were not screened out.

Breast Cancer Cohort **-** 200 women who were prescreened as positive for breast cancer by mammography and biopsy. The cohort represents a normal population of women presenting with breast cancer by cancer stage as well and women with stage 0 breast cancer**. In addition to achieving high (>95%) predictive power, the OTraces method correctly identified all stage 0 and stage 1 cancers in the samples with breast cancer**.

Prostate Cancer

The Prostate Cancer Test Panel

The Prostate cancer test panel described herein, *PC Sera Dx*, uses the same five biomarkers and demonstrated 90% predictive power when separating aggressive prostate cancer (Gleason score 7(4+3), 8, 9 and 10), from cancer-free samples. The ROC curve for this trial is shown in figure 7, below.

The blinded prostate cancer samples were measured at the Brady Urology Institute at the Johns Hopkins Medical Center (JHU) in Baltimore MD. The topmost (blue) curve shows the five-dimensional results using the OTraces method described herein. The can be compared to the green curve, which are the results from the well-known and currently approved PSA screening test for prostate cancer. The current PSA test achieves 90% sensitivity but sacrifices specificity which is only 57%. The OTraces method described herein achieves 95% sensitivity with a specificity of about 90%, substantially better than the PSA test. (Note that adjustment of the OTraces ROC curve decision point to achieve 98% sensitivity will cause a drop to about 85% specificity, still substantially better than the current screening test).

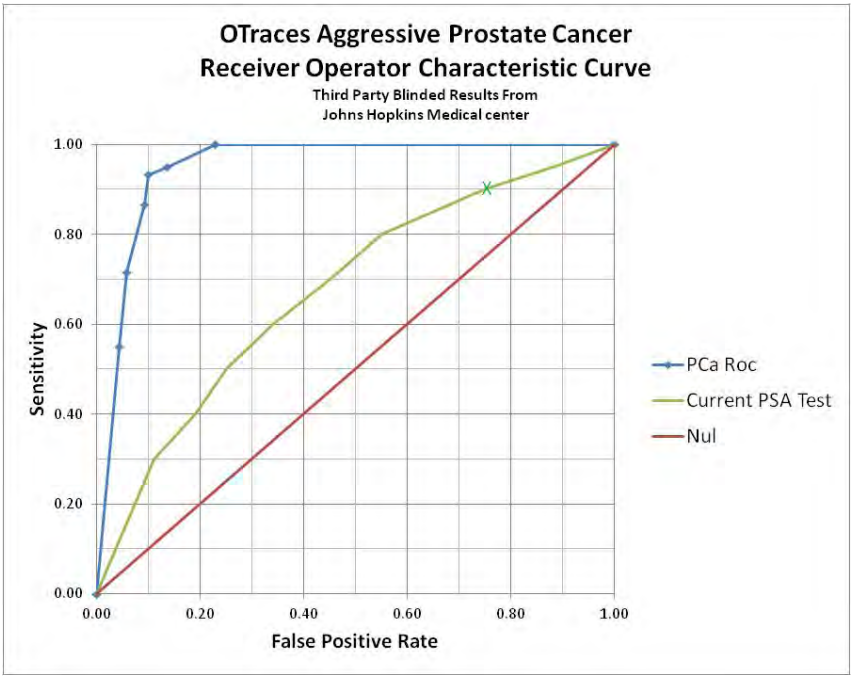

Figure 7. Prostate Cancer ROC curve

Cancer-free Cohort - This cohort (140 samples) had 20 samples with elevated PSA levels of up to 124 ng/ml (above the 4 ng/ml cutoff for medically required biopsy). **These samples were all called correctly as cancer-free by the OTraces method**. They simply showed no response for the biomarkers specific to the TME.

Prostate Cancer Cohort - The cancer-positive samples were from the JHU serum bank (60 samples). These samples were highly characterized and were aggressive, high Gleason score prostate cancers that would require immediate intervention.

RESOLVING TUMOR STATUS (AGGRESSIVENESS OR STAGE)

Utilizing information from the Tumor Microenvironment

We have found that the cytokines present in the TME have the information necessary to resolve tumor status, such as cancer stage or whether the tumor is quiescent (not growing) or not. Our research has shown that biomarker activity changes depending on the stage of the tumor. Early stage tumors strongly up-regulate the *IL8* protein (angiogenesis) to improve blood circulation in the surrounding tissue, before the tumor has amassed sufficient bulk to express *VEGF* to vascularize the tumor itself. At later stages this cytokine may subside and. *VEGF* up-regulates (tumor vascularization) to improve blood circulation within the bulk of the larger tumor. Early stage tumors may also show a strong pro-inflammatory response (*IL6*), which may subside as the later stage tumor successfully suppresses the immune system.

Figure 8 below shows the actions of four of these functional cytokines (and one tumor marker) for the same 400 breast cancer subjects (Gertsen Institute Study) as shown earlier. The graph shows average actions of *IL6, IL8, TNFα, VEGF* and a tumor marker (*Kallikrein 3* also known as PSA) plotted by cancer stage for these 400 subjects.

On the horizontal axis, -1 is women without cancer (putatively healthy). The stage (as determined by biopsy for those subjects with cancer) is plotted as 0, 1, 2, and 3 are also plotted on the horizontal axis in order with the -1 samples. As can be seen, three of these biomarkers sharply up-regulate at stage 0 breast cancer then subside as the tumors increase in stage. *VEGF* up-regulates in progression with stage as the tumor grows. **The average cancer score for the cancer stage is shown as the top line (tan). It progresses from an average of 20 for the cancer-free subjects and spikes up to an average of 190 for stage 0.** (The cancer scoring is an arbitrary range of 0 to 200 with 100 being the midpoint and 100 to 200 being cancer positive.) A predictive model can be constructed from this data using the same techniques as discussed above.

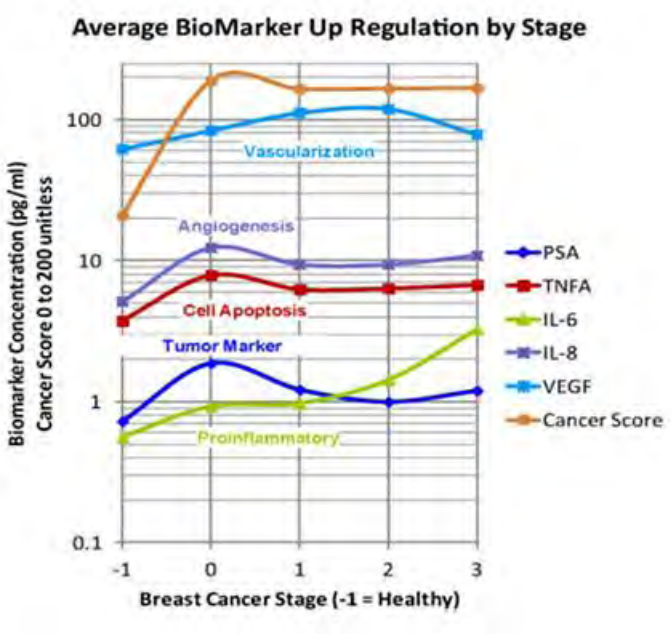

Figure 8. Biomarker Surge by Stage

Such a model was constructed using 190 of the breast cancer samples[13] from the Gertsen Institute that predicted the stage of cancer for each subject. Using these TME-derived proteins and the methods described above, the model demonstrated 95%+ accuracy. Though treatment protocol must ultimately be determined by biopsy, this model can determine if a subject has breast cancer and what stage it is at. The data suggest that this method may be able to "see" the tumor before it shows up in imaging. It should be noted that historically, the number of stage 0 breast cancers (which tend to be 2mm or smaller) detected by imaging (mammogram) is a small fraction of those found with stage 1 or higher. As a follow-on study, the Gertsen Institute plans to start recording and monitoring those patients who present with cancer and whom the OTraces method identified as stage 0, but whose tumor cannot be seen in imaging.

Similar results have been observed with prostate and lung cancer studies. In the case of prostate cancer, the data suggests that monitoring of the proteins in the TME proxy – serum - can determine if the tumor is quiescent, low Gleason Score, or high Gleason Score (growing aggressively and requiring intervention).

SUMMARY

The OTraces method described above involves six patented steps summarized below. The process involves building the TME based model (steps 1 through 6) (using a training set protocol) which creates the serum-based TME proxy. This is a multidimensional grid model where the number of dimensions is equal to the number of biomarkers being evaluated. Unknown samples are then scored by plotting them onto this grid. The score is determined by proximity to the training set data points. Those determined to be cancerous are in proximity to those training set data points which are mostly determined to be cancer, with the cancer-free samples are in proximity to the cancer-free training set data points. The steps are:

1. *Biomarker Selection* - Focusing on the TME and TIF as a source, selecting proteins active in the competition between the immune system to suppress the tumor and the tumor actions as it strives to survive and grow for a particular cancer type. These proteins should not overlap in functionality.
2. *Training Set Construction – Cancer-Free (Cohort A)* - Determine the mean values from the proteins in serum for a large sample population that does not have a TME for the tumor of interest. This population must include various non-cancer maladies that affect a normal population at a statistically significant level such that it represents a general population cohort that would present for an annual cancer screening test.
3. *Training Set Construction – Cancerous (Cohort B)* - Determine the biomarker mean values from the serum for a large sample population that has a known cancer of interest. Include in this population all non-malignant conditions in the percentage they normally exist in the population that present for screening.
4. Determine the mean values of each biomarker in cohort A and B.

[13] Stage data was only supplied for 190 of the 200 samples.

5. *Age normalization* - Correct for the age-related drift in the protein mean values.
6. *Noise Suppression* - Mathematically suppress the noise by comparing the unknown samples to a training set with these mean values as the anchor points, such that the training set is substantially correct.
7. *Proximity Scoring* - Plot the resulting new values for the concentration based proximity scores in multidimensional space.

This constitutes the TME based model. Once completed, unknown samples can be plotted onto this multidimensional grid and scored by proximity to the training set points of each type, either cancer-free or not.

The two biomarker mean values produced in steps 2 and 3 combined with the age normalization (step 4) and noise suppression (step 5) constitute the serum proxy for the TME-active proteins. Steps 4 through 6 are accomplished with the OTraces-developed proprietary algorithm.

The differences in the normalized and noise suppressed mean values are the “signal” or “signature” for the TME and thus the cancer. Focusing on the TME signature will eliminate identification of pre-cancerous conditions where only tumor related mutated DNA is present, but an active, life threatening tumor does not yet exist.

The process can be repeated using the same or other variables to enable tracking of tumor behavior and severity by replacing the known disease-free and disease state samples with those identified, for example, as "aggressive" and "non-aggressive" (or early-stage) tumors and repeating the evaluation steps.