How one thousand proteins in almost one thousand subjects help us stratify cardiovascular risk

Bob Mehler1, Alex Stewart1, Shintaro Kato3, Richard Lawn1, Steve Williams1, Christine Wells2, Marc Segal2, Peter Ganz4

Somalogic, Boulder CO1, UCSF2, NEC Corporation of America, Boulder, CO3, San Francisco General Hospital and University of California4

Here we demonstrate the power of applying high dimensional proteomics to the well-studied field of cardiovascular secondary event risk stratification.

Proteomic biomarker discovery in cardiovascular clinical studies is often biased by the selection of a few promising proteins already strongly suspected to be involved in the disease. This bias prevents the discovery of truly novel markers without any obvious connection. A multiplex proteomics assay (SomaScan®) developed by Somalogic allowed us to perform an unbiased biomarker search of over a thousand proteins, akin to using a wide angle lens as opposed to looking through the keyhole at a few known or suspected existing protein markers.

Over 1000 proteins were assayed from a single small blood sample drawn from each of 987 Individuals with stable coronary heart disease and a median 6 years of follow-up. Using an L1 regularized technique (LASSO) we built a linear Cox proportional hazard model using 11 proteins. The hazard ratios for the 11 Individual biomarkers range up to 5 between the highest and lowest quartiles for all subsequent CV events (myocardial infarction, heart failure, stroke, transient ischemic attack and mortality over eight years), compared to just 2.4 for the commonly-used biomarker hsCRP.

Compared to the existing Framingham secondary event model applied to this cohort we found that adding the proteomic information substantially improved the ROC; whereas the converse was not true. The proteomic model was also much more predictive over shorter (two year) timescales than the Framingham secondary event model.

The rich data sets generated also yield information about clinical sub-groups of the 987 Individuals in the study. We show how the strength of some markers in this study depends upon the personal context, as indicated by individual factors such as gender, age and diabetes. Such protein markers interpreted in the light of the specific individual enable the development of truly personalized diagnostic tests.

This study demonstrates how a single high dimensional proteomic study of a thousand subjects can radically expand knowledge in an established field, and may enable targeting of preventive therapies, enhance patient management, enrich enrollment for CV events in clinical trials, and identify potential targets for therapeutic discovery.

SOMAscan: Enabling Proteomic Discovery

Precise quantification using SOMAmerTM (Slow Off-Rate Modified Aptamers) reagents
- Sensitive and precise: LOD <1pM with <5% CV
- Highly multiplexed platform: >1000 analytes from 20 μl sample
- High throughput: >300 samples/day
- Broad coverage of cellular pathways and disease pathophysiology

Results – Strong novel protein biomarkers discovered

Most of these proteins are novel for the field

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Concentration</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angiopoietin-2</td>
<td>High</td>
<td>1.5</td>
</tr>
<tr>
<td>MMP-1</td>
<td>Low</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The wide aperture SomaScanTM view of blood proteomics reveals several strong independently prognostic markers of future cardiovascular events. Although some of these were known from other studies and conditions, only here can they be simultaneously compared and evaluated for common and independent information content. The co-variation is as relevant here as the individual predictive power. The Lasso statistical technique allows us to select markers which work well in concert.

Results-Comparison with current methods

Framingham model quartiles

Event free survival

On the right the ROCs demonstrate how the protein model outperforms the Framingham model. The combination of the two models only marginaly improves upon the protein model alone.

Framingham model includes Age, Cholesterol (HDL/LDL), Smoking, Gender, Diabetes, Blood pressure.

Context specific markers for personalized CV risk assessment

The performance of the general proteomic CV event-risk model can be improved by supplementing the model with proteins specific to relevant population subgroups. Such techniques can be used to personalize models of CV event risk.

Proteomic CV risk model:

Event free survival by quartile

A simple log-linear cox proportional hazard model is used to combine the eleven markers into a single hazard model. This model has particularly strong performance over the next two years, as might be expected for a proteomic measurement, responding to the actual state of the physiology, as opposed to a genetic model which allows for a potential propensity. The combination of Proteomic markers within a genetic context remains to be explored, although the work below on individually specific diagnostics indicates this may become a fruitful approach.

Summary

There are biologic differences among patients with known CHD, and some of those differences determine the likelihood of subsequent CV events. By recognizing those that are influential we can look forward to a future when we are beyond assuming uniform risk and beyond assigning uniform secondary prevention and risk reduction therapy for these patients. This powerful new tool may lead to:

- more cost-effective patient management through targeted treatment of the higher-risk populations
- improvement of secondary prevention clinical trial outcomes through enrichment of the target population
- assessment of relevant efficacy or toxicity response to drug treatment protocols
- identification of potential new targets for therapeutic discovery and development.

Statistical Methods Used to Construct Classifier

Goal: Identify top proteins which best predict recurrent CV events L1 penalized “Lasso” Cox PH models. Create a “risk score” for prediction of recurrent events

Ten-fold cross-validation
- Derives models from 10% of observations and validates them in the remaining 90% of observations
- This is repeated 10 times (for each 10%) to identify most consistent predictors
- LASSO = “Least Absolute Shrinkage and Selection Operator” Of 1000 proteins Lasso Finds the aptamer that is most strongly associated with survival, then applies shrinkage and cross-validates this model. Then Lasso finds the second aptamer to add it to the model and again shrinks both coefficients and cross-validates the model, and so on...

Advantageous model when the number of predictors far outnumbers the number of outcomes (microarrays)

Provides a natural way to encourage sparsity and simplicity in the solution