Efficiently generating multi-biomarker ROC curves to identify significant multi-biomarkers

Amol Prakash[1], Michael Athanas[2], David Sarracino[1], Bryan Krastins[1], Taha Rezai[1], Mary Lopez[1]
[1]Thermo Fisher Scientific, Cambridge, MA, USA [2]Vast Scientific, Wayland, MA, USA

Overview

Purpose: To demonstrate an efficient algorithm for combining multiple marker candidates to build a panel of candidate biomarkers.

Methods: A new algorithm is presented and is tested on simulated data showing expression ratios from 1000s of proteins in 100s of samples.

Results: ROC analysis in discovery is a robust discriminate in cohort studies. Efficiently combining multiple markers in some circumstances may provide a more revealing discriminate.

Introduction

Proteomic discovery experiments are rapidly generating lists of putative biomarkers for diseases and pathologies. Verification of these markers in multiplexed assays poses a statistical challenge as traditional ROC (Receiver Operation Characteristic) curves used to calculate the sensitivity and specificity of a diagnostic or predictive assay are based on single markers. The ability to combine quantitative information from several markers could potentially improve the diagnostic accuracy of existing tests and facilitate the development of new tests. However, standardized approaches to representing panels of markers remains controversial. However multi-marker ROC curves are computationally expensive to compute, and therefore have not been used. We suggest a novel algorithm to enable efficient computation of ROC curves of pair-markers and apply it to a large data set to identify significant pair markers.

ROC for Discovery?

A ROC curve is a graphical representation of the accuracy of a test to discriminate two classes (disease versus normal). It is simply a plot of the true positive (TP) rate as a function of the corresponding false positive (FP) rate as the discriminate threshold is varied (Figure 1). In the context of a study with multiple patient measurements of various characteristics (putative biomarkers), the TP and FP are calculated by accumulating the number of normal (disease) patients with above (below) a given threshold of a measurement. A plot of FP as a function of TP is then constructed by scanning the threshold over all relevant values.

The beauty of a ROC curve is that the resolving power of a given putative marker to distinguish two classes (normal vs disease) can be expressed as a single number: the area under the ROC curve. ROC plots are normalized so that the maximum area is 1.0. ROC curves with Area Under the Curve (AUC) close to 1 have high selectivity and sensitivity; whereas, curves with areas close to 0.5 correspond to cases where the putative marker effectively cannot distinguish the two classes.

ROC curves are readily used in clinical study reports as a concise visual representation describing the outcome of a specific test measurement on a patient population. In contrast, relative fold-change or expression ratio are used as discriminates in discovery experiments despite the fact that the actual ratio or fold-change is not as relevant as the marker's ability to distinguish normal from disease cases.

Figure 1. a traditional ROC plot is constructed by tabulating the FPs and TPs as the criteria threshold is swept across both of the curves.

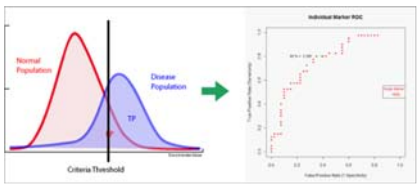
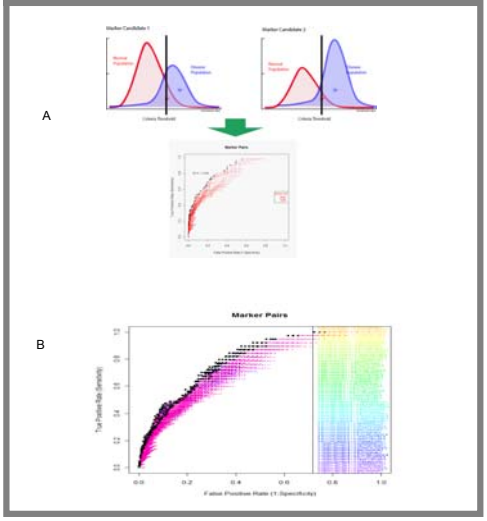


FIGURE 2. (A) Plot of a pair-marker ROC curve, where the different thresholds are varied for the two different marker populations resulting in the red points in the plot, and the black dots (highest true positive rate for a given false positive rate) generate the final ROC curve. (B) A ROC scatter plot of all possible pairs of the top 10 candidate peptide markers. The color points represent a possible FP and TP threshold value for a pair of peptides (legend on the right). The black points represent the outer edge of each individual ROC scatter plot. The outer edge represents the optimal ROC for each pair



Multi-Marker Panels

In a biological system, there may be multiple marker candidates that work in tandem with their own individual discriminating capability. The overall discriminating capability may be improved if a panel of markers were used instead of a single marker. Two or more marker candidates can be combined using the Marker Multiplex ROC described below. In this case, the TP and FP is derived by the combined probability for each criteria threshold for each marker, i.e., for a given set of N markers $m_1,m_2,\ldots, m_n$ and their criteria threshold $t_1,t_2,\ldots,t_n$, and for samples $s_1, s_2,\ldots,s_m$ and control $c_1, c_2,\ldots,c_m$

$$TP = \bigcap_{i=1}^{n} i\{\bigcup_{j} j\,|s_{ji} \geq t_i\} \qquad FP = \bigcap_{i=1}^{n} i\{\bigcup_{j} j\,|c_{ji} \geq t_i\}$$

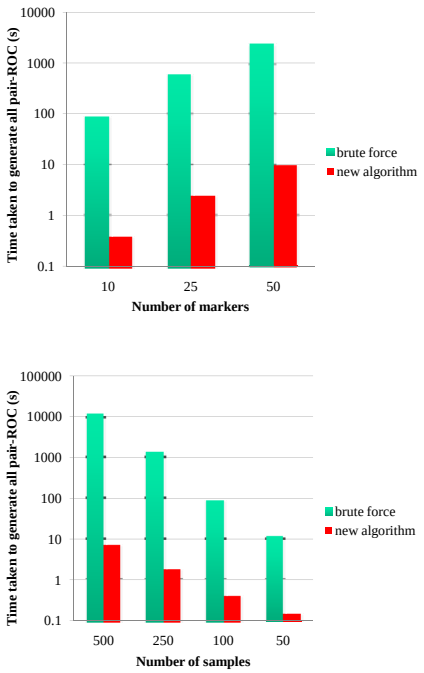
For the example illustrated above, a traditional ROC plot is constructed by tabulating the FPs and TPs as the criteria threshold is swept across both of the curves (Figure 2). The most effective discriminating power of the combined marker set is found as the top-leftmost edge of the scatter plot shown above as black points. An overall efficacy of the marker panel can be expressed as the area under the curve obtained by joining the points along the top-leftmost edge of the distribution. The most effective discriminating power of the combined marker set is found as the top-leftmost edge of the scatter plot shown above as black points.

Algorithm

Our optimizations are based on the fact that these thresholds for these black points can be calculated beforehand without considering every possible threshold pair. Suppose we want to find the sets of thresholds that yield x% false positive rate. Let the 100 controls be $c_1, c_2,\ldots,c_{100}$ so that these are sorted by the first marker, i.e., $c_{11} \leq c_{21} \leq \ldots \leq c_{100,1}$. Then:

Set of criteria = $\{\{c_{x1},\infty\}\}$
pointer = x; sorted_list=$\{c_{12}, c_{22},\ldots,c_{x2}\}$
repeat while pointer<100
  pointer++
  Add $c_{pointer,2}$to sorted_list
  Remove maximum value from sorted_list, to ensure size=x
  While ($c_{pointer,2}$≥maximum value of sorted list)
    Pointer++
  Add { $c_{pointer,1}$,maximum value of sorted list} to sets of criteria
For each criteria, find count of samples { $s_1, s_2,\ldots,s_m$} that pass that criteria
TPx = maximum count of samples amongst all criteria for false positive rate = x%

FIGURE 3. Performance of the new algorithm when compared to the brute force algorithm on simulated data. Both plots have logarithmic y-axis. In top panel, we vary the number of markers, keeping the number of samples constat=100. In the lower panel we vary the number of samples, keeping the number of makers constant=10.



Results

We simulated protein expression values for a number of samples and a number of proteins. Using this data set, we computed the ROC curves for every pair of markers, using the brute force algorithm (Figure 2) and the novel algorithm presented. Figure 3 plots the time benefits of the novel algorithm as we change the number of makers and samples.

Conclusion

We present a novel algorithm to efficiently compute ROC curves of pair-markers.

- ROC analysis at the discovery phase of an experiment is a powerful way of identifying robust candidate biomarkers.
- Multiple marker distinguishing capability can be significantly enhance when compared to single marker capability.
- A novel algorithm helps compute the ROC curve in a computationally feasible time.
- The significance of triplets and higher order marker combinations is diminished.

References

1. Discovery analysis of multiple protein marker panels: Optimizing sensitivity and selectivity. Athanas et al., MSACL 2010, San Diego, CA
2. M, Lopez, R. Kuppusamy, D. Sarracino, A. Prakash, M. Athanas, B. Krastins, T. Rezai, J. Sutton, S. Peterman, and K. Nicolaides; *Discovery and targeted SRM assay development of first-trimester peptide biomarker candidates for Trisomy 21 in maternal blood*, submitted for publication
3. Lukas Käll, Jesse Canterbury, Jason Weston, William Stafford Noble and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets Nature Methods 4:923 – 925, November 2007