Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multiple Survival Screening algorithm (MSS) for identifying high-quality cancer prognostic markers

Edwin Wang (edwin.wang@cnrc-nrc.gc.ca)
Wang's Lab, systems biology
Jie Li
Wang's Lab, systems biology

Method Article

Keywords: canccer, biomarker, prognosis, gene expression, algorithm

Posted Date: February 7th, 2011

DOI: https://doi.org/10.1038/protex.2011.211

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

We have developed a Multiple Survival Screening algorithm \(MSS) for identifying high-quality cancer prognostic markers from the gene expression profiles of cancer samples. By applying the MSS algorithm to breast cancer samples, we have identified several marker sets which showed ~90% predicting accuracy across 8 independent breast cancer cohorts. We realized that the algorithm could be used for finding other biomarkers including drug response markers. We are describing the protocol with some comments based on our experience in using the algorithm.

Procedure

The MSS algorithm includes 9 steps. Each step has several parameters. In each step, we will give our suggestions for the parameters. The details of the terms and parameters mentioned in the protocol can be found in the original article \(Li et al., Nature Communications, 1:34, 2010). Step 1 A survival gene pool is generated in this step by performing a genome-wide single gene survival analysis in a given training dataset. Generally the survival p-value is less than 0.05. To get more robust markers, we suggest that several such training datasets could be used to obtain several survival gene pools. These gene pools could be merged for the next step. Step 2 In this step, cancer hallmark related Gene Ontology \(GO)-term-defined gene sets could be generated by functional annotation of the survival genes, which have been generated in Step 1, using GO analysis tools such as DAVID Bioinformatics Resources \(http://david.abcc.ncifcrf.gov/). Normally each of the GO-term-defined gene sets contains 50~100 genes.. If a GO-term-defined gene set includes not many genes \(i.g., less than 45), we usually discarded it. Alternatively, we combined the genes \(to get the gene size between 50 and 100 genes) from a few GO-term-defined gene sets in which each GO-term-defined gene set contains less than 50 genes. If a GO-term-defined gene set includes many genes \(i.g., more than 100), we ranked the genes by running the Steps 3-7 and took the top 60-80 genes for running the MSS. Step 3 Random gene sets \(RGSs) have been generated in this step. We generated 1 million distinct RGSs from each selected GO-term-defined gene set. Each RGS contains 30 genes. More RGSs could be generated if you have powerful computer clusters, thus the biomarker might be more robust. Step 4 Random datasets \(RDSs) have been generated in this step using the training dataset. We normally generated 36 RDSs. However, more RDSs could be generated when powerful computer clusters are available. It is critical to maintain the same ratio of “good” and “bad” tumors as that in the original training set. Furthermore, it is better to have at least 60 samples for each RDS. It is better to make sure that these RDSs have the maximal difference of the samples. Steps 5-7 These steps are used to perform survival screenings of the RGSs on the RDSs. Several parameters need to be set in these steps. In the MSS algorithm, we selected the “predictive RGSs” whose survival p-value is less than 0.05 in more than 90% of the RDSs \(RDS passing rate). However, we found the p-values could be less than 0.01, while the RDS passing rate could be between 75% and 95%. These parameters could be adjusted to have several thousand RGSs selected. The selected RGSs can be used to get the top 30 most frequent genes \(a potential gene signature). If more than 30% of the 1 million RGSs have p-values less than 0.05 in more than 35 RDSs, this experiment will be discarded. The results might be come from data

overfitting. On the other hand, if only a few hundred RGSs have p-values less than 0.05 in less than 80% of the RDSs, this experiment will also be discarded. In addition, it may need to run more RGSs \(4 or 8 millions) to get enough selected RGSs. This depends on the datasets and the GO-term-defined gene set. Step 8 A potential gene signature containing top-ranked 30 most frequent genes has been obtained from the selected RGSs. In fact, the range of the gene size of a potential gene signature could be between 20 and 30 based on the GO-term-defined gene contents and training datasets. Step 9 This step is to assess the reproducibility and stability of a potential gene signature. After running 2 distinct 1 million RGSs derived from one GO-term-defined gene set on the same RDS set, we examined how many genes are in common between the two top-ranked 30 genes. The common genes could be from 20 to 30 based on different experimental conditions. Furthermore, the top-ranked 30 genes may be different when different RDSs, RGSs, parameters and training data sets were used. However, the performance of the selected gene signatures should be robust in other independent testing datasets, for examples, they often have survival p-value \(<0.05) in the testing datasets if the MSS protocol has been followed.

Troubleshooting

Suggestions have been embedded in each step.

References

Li et al., Identification of high-quality cancer prognostic markers and metastasis network modules, Nature Commununications 1:34 \(2010) | doi: 10.1038/ncomms1033