Gene expression profiling robustly predicts the outcome of patients diagnosed with early stage lung adenocarcinoma

Yann Gaston-Mathé* [1], Charles Ferté[2], Benoit Gautier[1], Ndjido Ardo Bar[1], Mathilde Bateson[3], David Planchard[2], Benjamin Besse[2], Jean-Pierre Armand[2], Jean-Charles Soria[2]

(*): corresponding author (yann.gastonmathe@hypercube-research.com) (1): HyperCube Research SARL; (2): Gustave Roussy; (3): Institut HyperCube

Background

Although the management of metastatic lung adenocarcinoma has been profoundly modified by the identification of actionable molecular traits, decision making for early-stage lung adenocarcinoma still relies on the tumor stage (TNM) only.

Most of these patients recur within the 5 years after the tumor resection (Stage I: 30%, Stage II: 50%, Stage III: 70%), leading to their death by cancer in most cases.

To improve the prediction of probability of overall survival in these patients is critical, in order to better identify:

- The stage I-II patients at high risk of recurrence or death that could benefit of adjuvant therapy.
- The stage II-IIIA patients at low risk of recurrence or death that could be spared unnecessary treatment.

The recent availability of high-dimensional molecular data (gene expression data) for lung adenocarcinoma patients, simultaneously to the developments of novel algorithmic models – such as the Hypercube® method - are expected to dramatically improve such predictive challenges.

In this work, we used publically available data sets to investigate the capacity of HyperCube, an innovative sub-group discovery algorithm, to discover robust predictive signatures of outcome.

HyperCube®

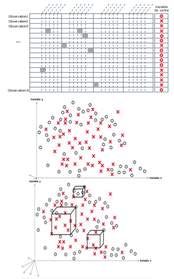
HyperCube (Hypercube Research, Paris, France, a subsidiary of BearingPoint, www.hypercube-research.com) is a novel data mining approach, that searches for local areas of over-density of a specific outcome in m-dimensional data space and outputs combinations of factors associated with high frequency of the studied outcome (rules). HyperCube technology enables to identify local interactions between variables impacting the outcome which are difficult to detect with classical statistical modeling or other machine learning algorithms.

HyperCube has been successfully applied in epidemiologic research (malaria), but until now remained unproven for the analysis of high dimensionality molecular data.

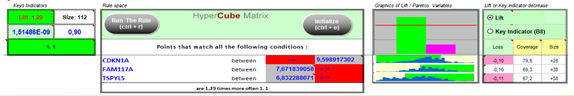
1. Data set with m input variables and 1 binary output variable is represented as m-dimensional space with observations as points of either modality

2. The algorithm randomly search for hypercubes with high concentration of points of the target modality

3. Each hypercube is optimized and filtered through 3 steps
   - Variable selection
   - Optimal boundaries selection
   - Rule generation



4. **Example of HyperCube rule:**



Methods

1. Data sets and data preparation

Inclusion criteria for patient selection:

- Publically available gene expression datasets
- Stage IA-3A Lung adenocarcinoma patients
- Surgical resection (R0) , no adjuvant therapy (CT, RT), Overall survival data (36M+ follow-up)

Four datasets were used:

- Directors Challenge (n=274) for rule generation, variable selection and model training;
- Hou et al (n=35) as "validation set" for variable selection;
- Zhu et al (JBR.10 trial) (n=31) and Rousseaux et al (n=86) as test data sets.

Gene expression data were normalized using RMA method and rescaled. 2540 genes were selected using Guinney's method. Clinical covariates were Stage, Grade, Age and Sex. The output variable was the 3-year survival (yes/no).

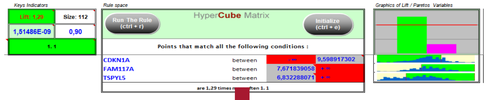
2. HyperCube learnings and rules selection

HyperCube was used to generate rules (>90% purity) on both modalities.

Rules were filtered to select only "open rules" (gene expression > threshold or < threshold) with purity >80% (rules for survival) and >70% (rules for death).

After filtering, 149 rules remained (80 for survival, 69 for death) and were used for subsequent variable selection.

3. Variable selection (for each valid rule – see example below):



Logistic Regression model built on training set (Dir) and applied on "test" set (Hou):

- **step 1:** os3yr ~ CDKN1A + FAM117A + TSPYL5
- **Step 2:** os3yr ~ CDKN1A + FAM117A + TSPYL5 + CDKN1A * FAM117A + CDKN1A * TSPYL5 + TSPYL5 * FAM117A
- **Step 3:** os3yr ~ CDKN1A + FAM117A + TSPYL5 + CDKN1A * FAM117A + CDKN1A * TSPYL5 + TSPYL5 * FAM117A + CDKN1A * FAM117A * TSPYL5

Performance test: AUC > 0.75 in training set (Dir) AND AUC > 0.7 in « test » set (Hou)

No → Variables removed

Yes → Variables selected for final model

4. Final model selection and training

Selected variables were included in an Elastic Net model cross-validated on the training set (Dir). Non-significant variables were removed.

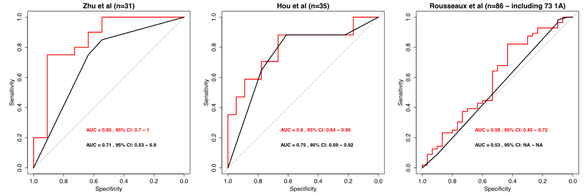
Final model trained on Dir, applied to 3 test data sets (Zhu, Hou, Rousseaux)

Results

A model comprising 9 genes + the Tumor Staging was developed. The model also comprised 3 2x2 interactions variables.

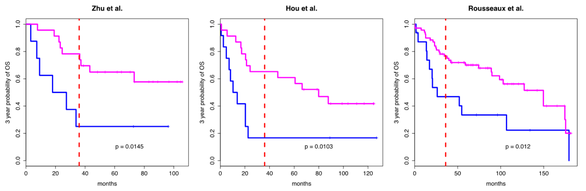
The model displayed a 0.8 AUC in the training set ROC. The model was robust in 3 distinct test data sets.

Model performance in 3 datasets measured by ROC AUC:



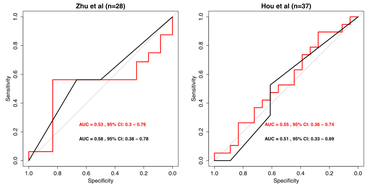
- Model performance consistently superior to Tumor Stage alone
- Small number of patients in validation cohorts impact on statistical significance vs Tumor Staging (developed on >10K pts from the SEER databases)
- Model performance in Rousseaux is lower possibly due to different composition of population (higher proportion of Stage 1 patients)

Kaplan Meier curves of overall survival for high- vs. low-risk predicted by model:



NB: High- and low- risk are defined by the cut-off of returned probabilities = 0.5

The model developed was specific to adenocarcinoma. When applied to LCC and SCC contingents in Hou and Zhu, the performance dropped.



Discussion

1. An innovative approach

This variable selection from a high-dimensionality gene expression data set produced a high- performing and robust prognosis classifier in 4 datasets where other methods have failed to achieve significant results compared to clinical staging performance (not disclosed here: ElasticNet, Random Forest, PLS, etc.)

This innovative approach relies on:

1. Identification of combined gene expression patterns associated to a high frequency of death/survival (rules)
2. Building simple RegLog models using the genes from the rules and interactions variables, and testing the models in both the training set and the "pseudo-validation" data set
3. Using only selected variables from step 2 to build the final model

We believe this approach may hold promises for future discoveries of meaningful and robust classifiers in high-dimensionality data sets.

2. Biological significance

From the 9 genes used in our model (not disclosed here), 8 of them have been reported to be linked with key oncogenic processes. Alterations have been reported for all 9 genes in TCGA lung adenocarcinoma patients (n= 129). Genes were altered in 31% of cases and alterations tended to be mutually exclusive.

The fact that the model is robust in 3 lung adenocarcinoma data sets but fails to accurately predict the outcome in LCC & SCC patients is indicative of an underlying biological heterogeneity and validates our choice to restrict our population to one tumor type only.

3. Limits

The current works suffers from two major limitations:

- Small number of patients impact on statistical significance
- The model still relies significantly on the Tumor Stage variable

Works are ongoing to identify new datasets for robustness assessment and models independent on tumor stage.

References

Ferté C, Trister AD, Huang E, Bot BM, Guinney J, Commo F, Sieberts S, André F, Besse B, Soria JC, Friend SH. Impact of bioinformatic procedures in the development and translation of high-throughput molecular classifiers in oncology. Clin Cancer Res. 2013 Aug 15;19(16):4315-25.

Loucoubar C, Paul R, Bar-Hen A, Huret A, Tall A, Sokhna C, Trape JF, Ly AB, Faye J, Badiane A, Diakhaby G, Sarr FD, Diop A, Sakuntabhai A, Bureau JF. An exhaustive, non-euclidean, non-parametric data mining tool for unraveling the complexity of biological systems--novel insights into malaria. PLoS One. 2011;6(9):e24085.

Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden K, Taylor JMG, Enkemann SA, Tsao M-S, Yeatman TJ, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med. 2008;14:822–7.

Zhu C-Q, Ding K, Strumpf D, Weir B a, Meyerson M, Pennell N, et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. J Clin Oncol. 2010;28:4417–24.

Hou J, Aerts J, Den Hamer B, Van Ijcken W, Den Bakker M, Riegman P, et al. Gene expression based classification of non-small cell lung carcinomas and survival prediction. PloS One. 2010;5:e10312.

Rousseaux S, Debernardi A, Jacquiau B, Vitte AL, Vesin A, Nagy-Mignotte H, Moro-Sibilot D, Brichon PY, Lantuejoul S, Hainaut P, Laffaire J, de Reyniès A, Beer DG, Timsit JF, Brambilla C, Brambilla E, Khochbin S. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Sci Transl Med. 2013 May 22;5(186):186ra66.

AACR ANNUAL MEETING 2014

April 5-9, 2014 • San Diego, CA

www.AACR.org • #AACR14