Comparative study on feature selection methods in ovarian cancer diagnosis using chemometric tools

Hae Woo Lee[1], Jay Liu[2], Young Jung Na[3] , Seongkyu Yoon[1],
[1]University of Massachusetts Lowe, MA; [2]Pukyong National University, Busan, Korea, Republic of; [3]CHA University, Seoul, Korea, Republic of

BACKGROUND

Ovarian cancer has the poorest prognosis of the gynecological cancers. 75% of the patients are diagnosed with stage III-IV on their diagnosis. The 5 year survival rate of ovarian cancer patients, whose disease was diagnosed at stage III or IV is less than 20%, but that of the patients diagnosed with stage I is higher than 90%.The early detection and diagnosis of cancer is one of the major application areas in omics technology and has received tremendous attention in last few years.

Bioinformatics plays a significant role especially for the analysis of high-dimensional, complex, and noisy omics datasets. The common difficulty in the feature selection for cancer-specific biomarker discovery is the inconsistency of the feature sets selected among different samples or by different selection methods. This issue is particularly crucial in biological applications where subsequent biomedical analysis of selected features requires relatively considerable time and costs.

RESEARCH TARGET

Problem definition

- Typical proteomic mass spectral data sets have higher number of variables (usually>10,000) than number of observation (usually<100) → curse of dimensionality
- To build statistically sound classification model from omics data sets, feature selection (biomarkers discovery) is a crucial step
- Many different feature selection methods, which have been reported until now, produce different selection results, which is difficult to be generalized or interpreted

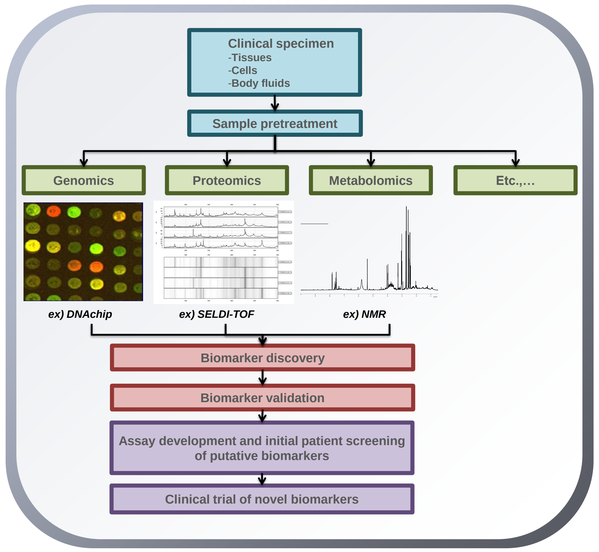
- **Confirm the feasibility of chemometrics tools in classifying cancer patients using omics profiles of human body fluids**
- **Suggest most suitable method for finding cancer-specific biomarkers on the basis of their consistency and accuracy**

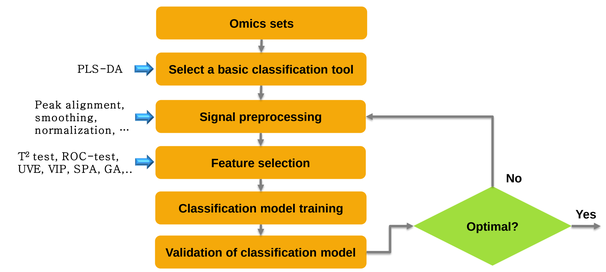
SUMMARY

1. Two data sets of MALDI-TOF-MS (proteomics) and LC/TOF-MS (metabolomics) spectra measured from serum samples to diagnose ovarian cancer were analyzed, revealing that some feature selection methods can suffer severely from the irreproducibility issue and this might deflate the potential benefits of omics technology for cancer diagnosis. In addition, the consistency of feature selection was highly dependent on the characteristics of dataset.

2. A possible remedy might be multi-objective formulation of the feature selection problem by simultaneously optimizing stability and accuracy of classification models, and this will be investigated further in future research.

APPROACH





General frameworks for biomarker discovery in clinical applications and detail flowcharts for bioinformatic analysis in analyzing omics dataset

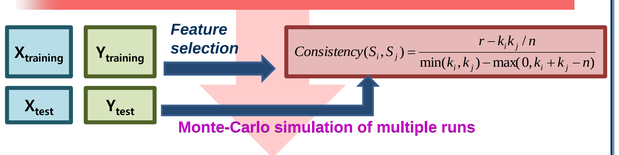
Feature selection methods employed in this study

Two-sample T^2 test | ROC-based method | Uninformative Variable Elimination (UVE) | Variable Importance in Projection (VIP) | Successive Projection Algorithm (SPA) | Genetic algorithm (GA)

Partial least squares-discriminant analysis (PLS-DA) algorithm

- **Modeling equaton:** $\mathbf{X} = \mathbf{TP}^T + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E}$ and $\mathbf{Y} = \mathbf{TC}^T + \mathbf{F} = \hat{\mathbf{Y}} + \mathbf{F}$
- **Classification rule:** $\hat{y}_{i,class} = 1$ if $\hat{y}_i \geq threshold$
  $\hat{y}_{i,class} = -1$ if $\hat{y}_i < threshold$

Consistency (reproducibility) of each feature selection methods were quantified by using Monte-Carlo simulation, where multiple runs of feature selection were conducted along with variations in the training dataset



CASE STUDY 1

LC/TOF-MS metabolomic data of human serum

Data sets

- LC/TOF-MS analysis of human serum samples obtained from patients with ovarian cancer [3].
- Positive and negative ion mode ESI spectra in the mass range of 100-1750 Da.
- Dataset was preprocessed, resulting in total 592 variables with 72 samples.

Table 1. Characteristics of Ovarian Cancer Patients and Controls

characteristics	Stage I/II	Stage III/IV	Controls	Total
Average Age (y)	60	61	54	58
Cancer	9	28	0	37
Control	0	0	35	35

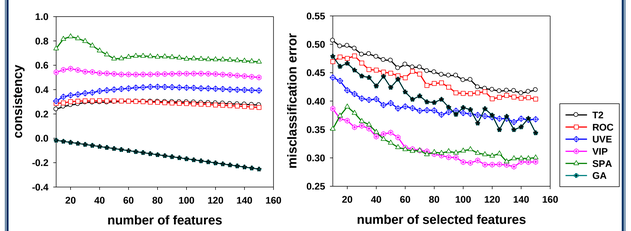


Fig 1. Comparison of different feature selection methods along with increasing number of selected features; In terms of their consistency and misclassification error, SPA and VIP gave superior results to other methods.

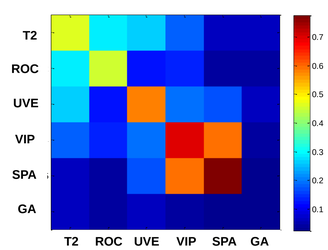


Fig 2. Consistency evaluated across different feature selection methods revealed that very distinctive features are selected among the methods. Due to the multivariate interaction exists in this dataset, VIP and SPA gave similar profiles.

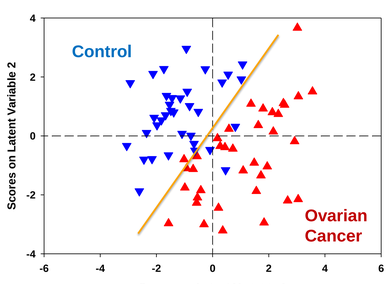


Fig 3. Example score plot of PLS-DA model build on the features selected by SPA method.

CASE STUDY 2

MALDI-MS proteomic data of human serum

Data sets

- MALDI-MS analysis of human serum samples obtained from patients with ovarian cancer [4].
- Data were preprocessed by smoothing, baseline removal, normalization and peak detection using Cromwell package and Bioinformatics toolbox in MATLAB, resulting in total 3438 m/z variables and 170 samples.

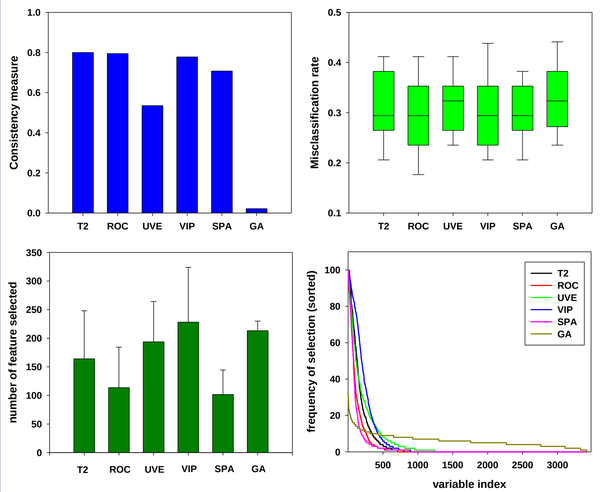


Fig 4. Comparison of different feature selection methods indicate that feature selection results were highly reproducible for this dataset, except for GA and UVE. In terms of overall efficiency and simplicity, ROC-based method gave best performance, followed with SPA.

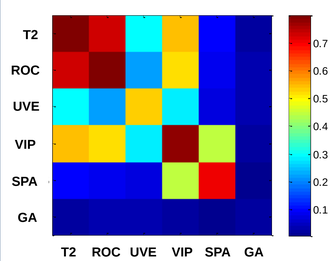


Fig 5. Consistency evaluated across different feature selection methods revealed similarity of selected feature profiles among different methods. Due to weak multivariate interaction, selected feature profiles were much similar than cast study 1.

REFERENCES

[1] Kalousis A, *et al. Knowl. Inf. Syst.* 2007;**12**:95-116.
[2] Lustgarten JL *et al.*, *AMIA Annu. Sym. Proc.* 2009;406-410.
[3] Guan W, *et al. BMC bioinformatics* 2009;**10**:259-263.
[4] Wu B, *et al. Cancer Inform.* 2006;**2**;123-132.
[5] Xiaobo X *et al.*, *Anal. Chim. Acta* 2010;**667**:14-32.