

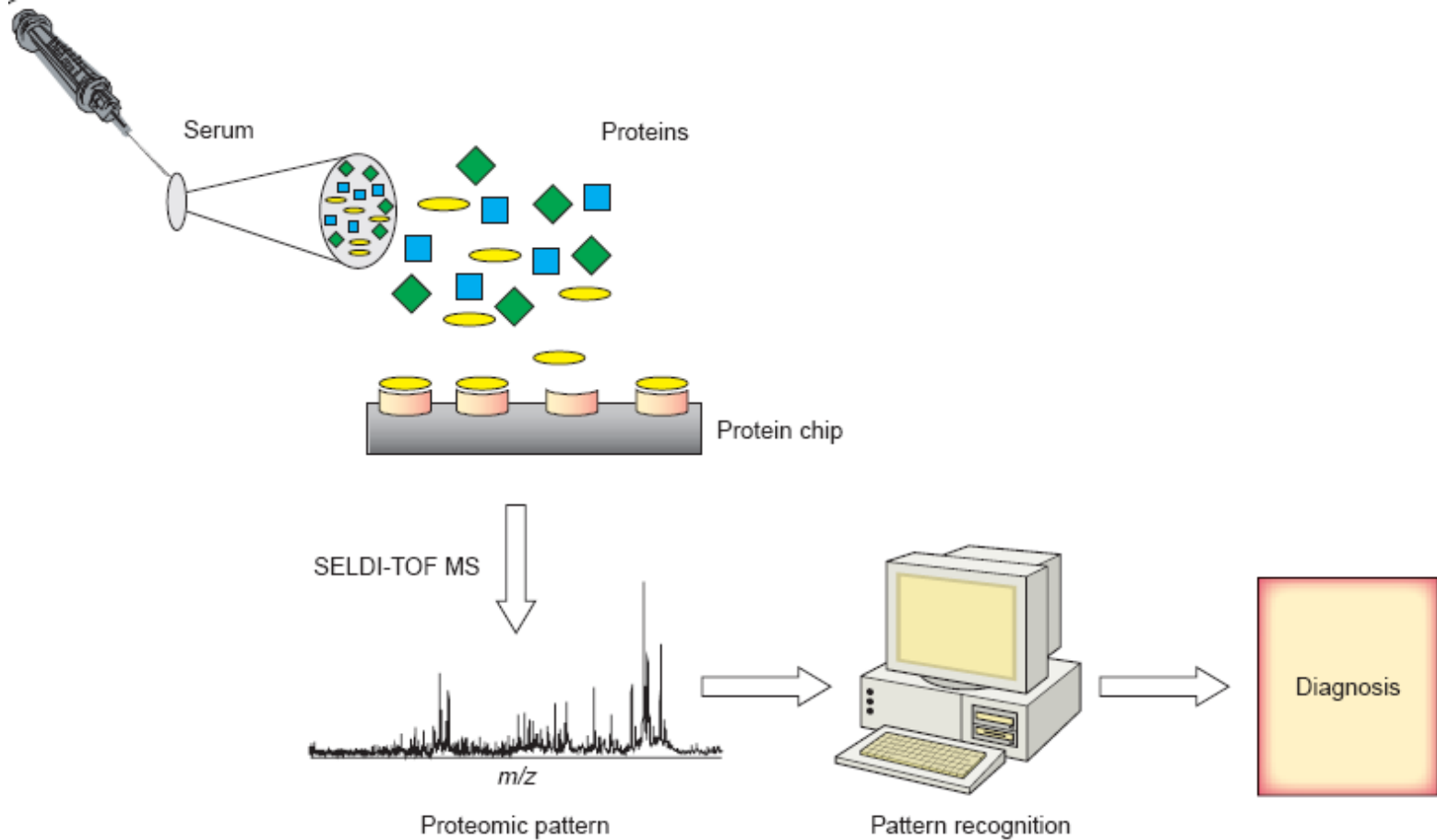
Dimensionality Reduction and Main Component Extraction of Mass Spectrometry Cancer Data

Yihui Liu

Introduction

Aim:

- Surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) in combination with advanced data mining algorithms, is used to detect protein patterns associated with diseases.



The sample obtained from the patient is applied to a protein chip which is made up of a specific chromatographic surface. After several washing steps and the application of an energy-absorbing molecule, the species that are retained on the surface of the chip are analyzed via mass spectrometry [1].

Dataset

Table 1. The features of three datasets.

Dataset		Dimension number	Sample number
High resolution ovarian	Cancer	368750	95
	Control	368750	121
Prostate cancer	Cancer	15154	69
	Control	15154	253
Low resolution ovarian	Cancer	15154	162
	Control	15154	91

Methods

Dimensionality reduction

- Feature Extraction:

PCA, LDA, KPCA, ICA

- Feature Selection:

Genetic algorithm (GA)

Sequential Forward Selection (SFS)

Our proposed method: Wavelet feature extraction

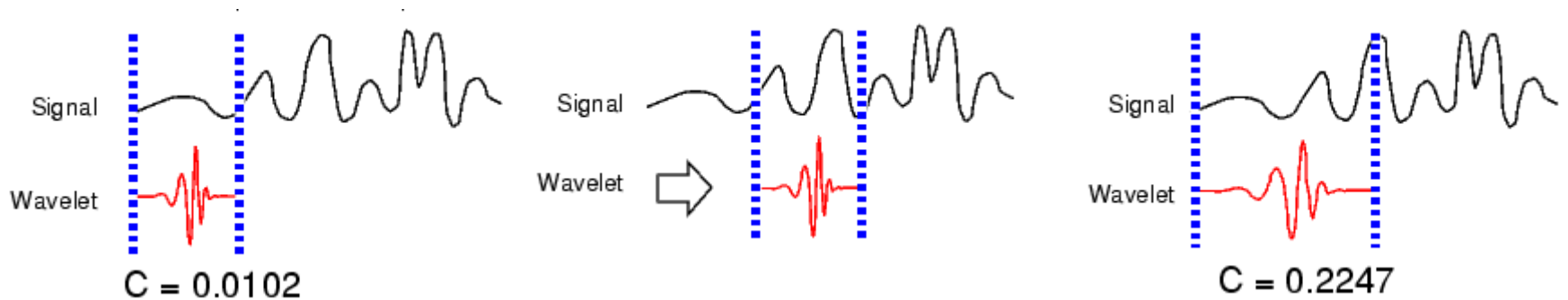
- wavelets tend to be irregular and asymmetric

A wavelet is a waveform of effectively limited duration that has an average value of zero.

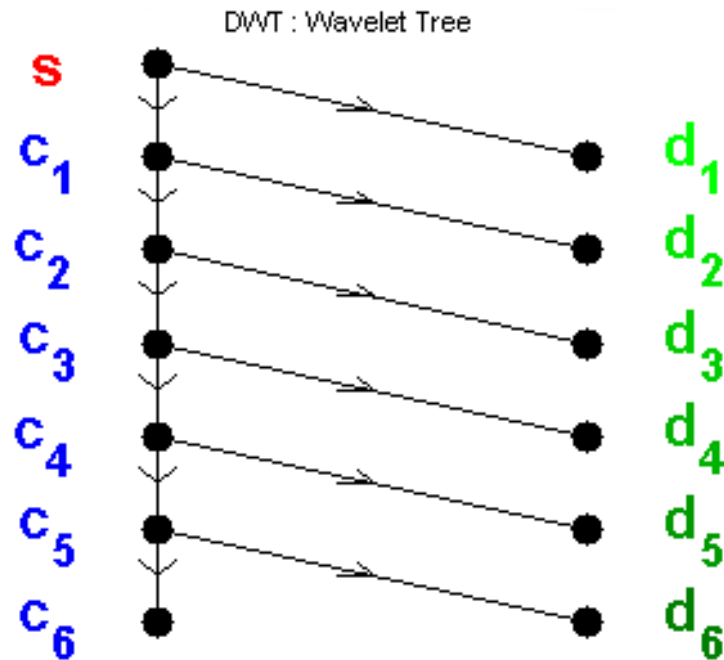


Wavelet (db10)

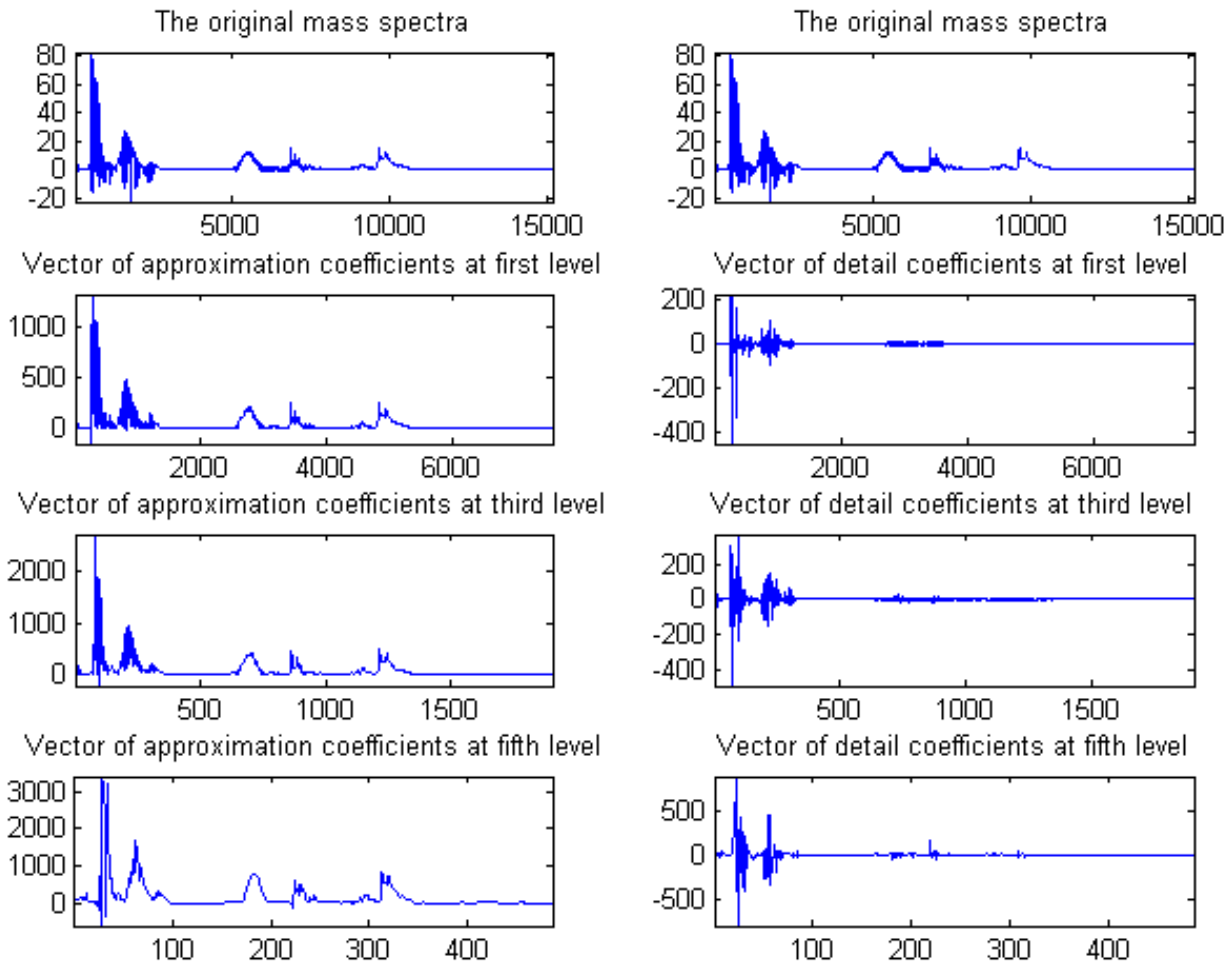
wavelet analysis is breaking up data vector into shifted and scaled versions of the original (or *mother*) wavelet.



$$C(\text{scale}, \text{position}) = \int_{-\infty}^{\infty} f(t)\psi(\text{scale}, \text{position}, t)dt$$



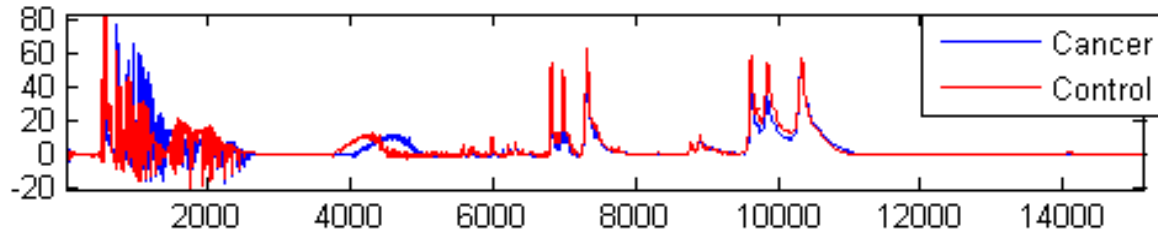
Multilevel wavelet decomposition tree for mass spectra. Symbol s represents mass spectra; c_i represent wavelet approximation coefficients from first level to sixth level respectively; d_i represent wavelet detail coefficients from first level to sixth level.



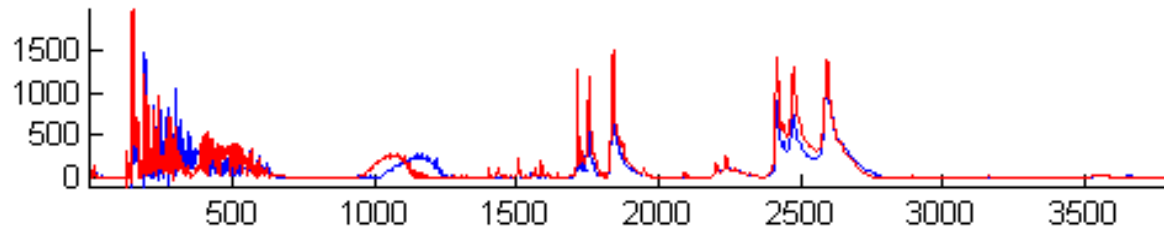
This figure show approximation and detail coefficients at first, third, fifth level. The dimensionality of approximation coefficients is 7583, 1905, and 486 for first level, third level, and fifth level.

Prostate cancer dataset

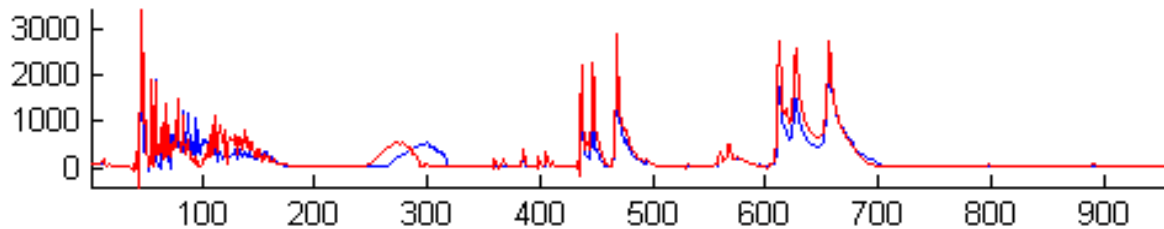
Vector of original mass spectra for prostate dataset



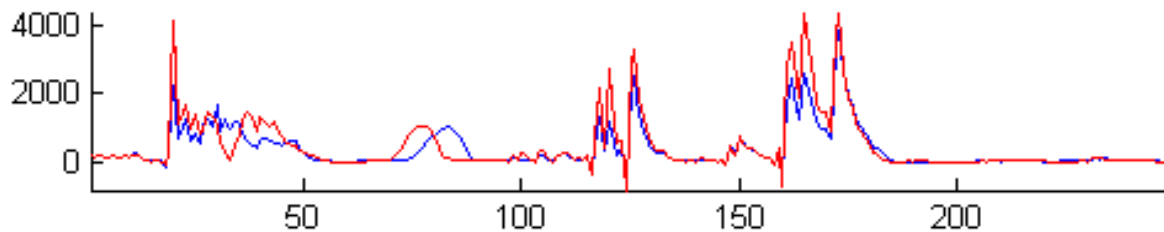
Vector of approximation coefficients at second level



Vector of approximation coefficients at fourth level



Vector of approximation coefficients at sixth level



This Table shows energy distribution of approximation coefficients for prostate cancer dataset of 322 samples. The fourth level is suggested as the best decomposition level in which approximation coefficients hold around 90% energy.

Approximation coefficients	Level 1		Level 2		Level 3		Level 4		Level 5		Level 6	
Dimensionality	7583		3798		1905		959		486		249	
Energy distribution (percentage)	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
	98.97	0.76	97.16	2.29	95.55	3.58	93.23	5.26	89.35	6.67	81.29	9.21

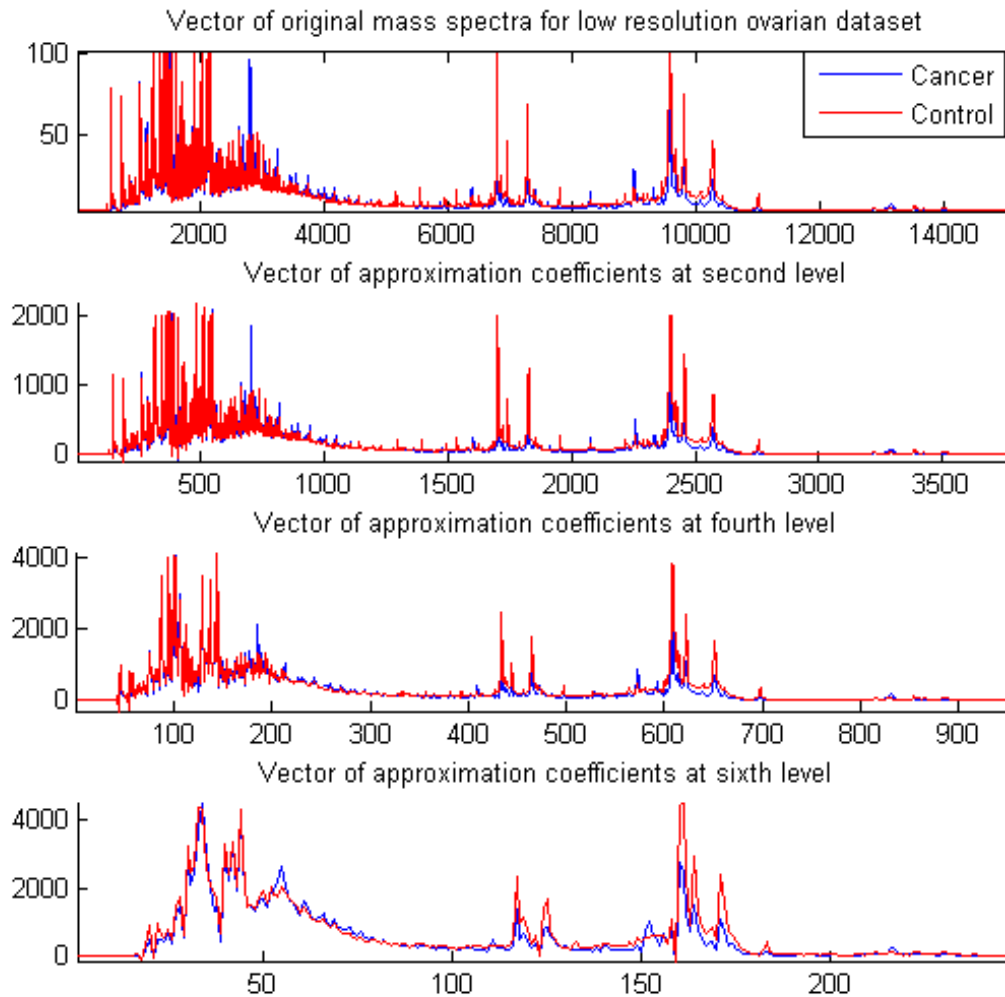
Method	level	<i>Correct rate</i>	Sensitivity	Specificity	PPV	NPV	BACC
Approximation coefficients	1	<i>0.9278</i>	0.9640	0.7948	0.9451	0.8577	0.8794
	2	<i>0.9282</i>	0.9654	0.7915	0.9444	0.8620	0.8785
	3	<i>0.9243</i>	0.9598	0.7938	0.9447	0.8435	0.8768
	4	<i>0.9213</i>	0.9570	0.7901	0.9436	0.8337	0.8736
	5	<i>0.9130</i>	0.9503	0.7765	0.9397	0.8098	0.8634
	6	<i>0.9048</i>	0.9417	0.7695	0.9374	0.7827	0.8556
Detail coefficients [9]	3						0.8618

This Table shows the performance for prostate cancer dataset based on three fold cross validation experiments, running fifty times (50x3CV).

Performance of different methods [2]

Methods	BACC	Specificity	Sensitivity	PPV
No FE	0.777	0.698	0.855	0.439
PCA	0.516	0.540	0.493	0.248
PCA/LDA	0.667	0.710	0.623	0.431
SFS	0.827	0.929	0.725	0.728
SBS	0.729	0.806	0.652	0.498
P-test	0.728	0.877	0.580	0.572
T-test	0.709	0.897	0.522	0.575
KS-test	0.784	0.857	0.710	0.579
NSC(20)	0.736	0.833	0.638	0.529
Boosted	0.810	0.881	0.739	0.627
Boosted FE	0.906	1.000	0.812	1.000

Low resolution ovarian dataset



Performance of approximation coefficients at six levels.

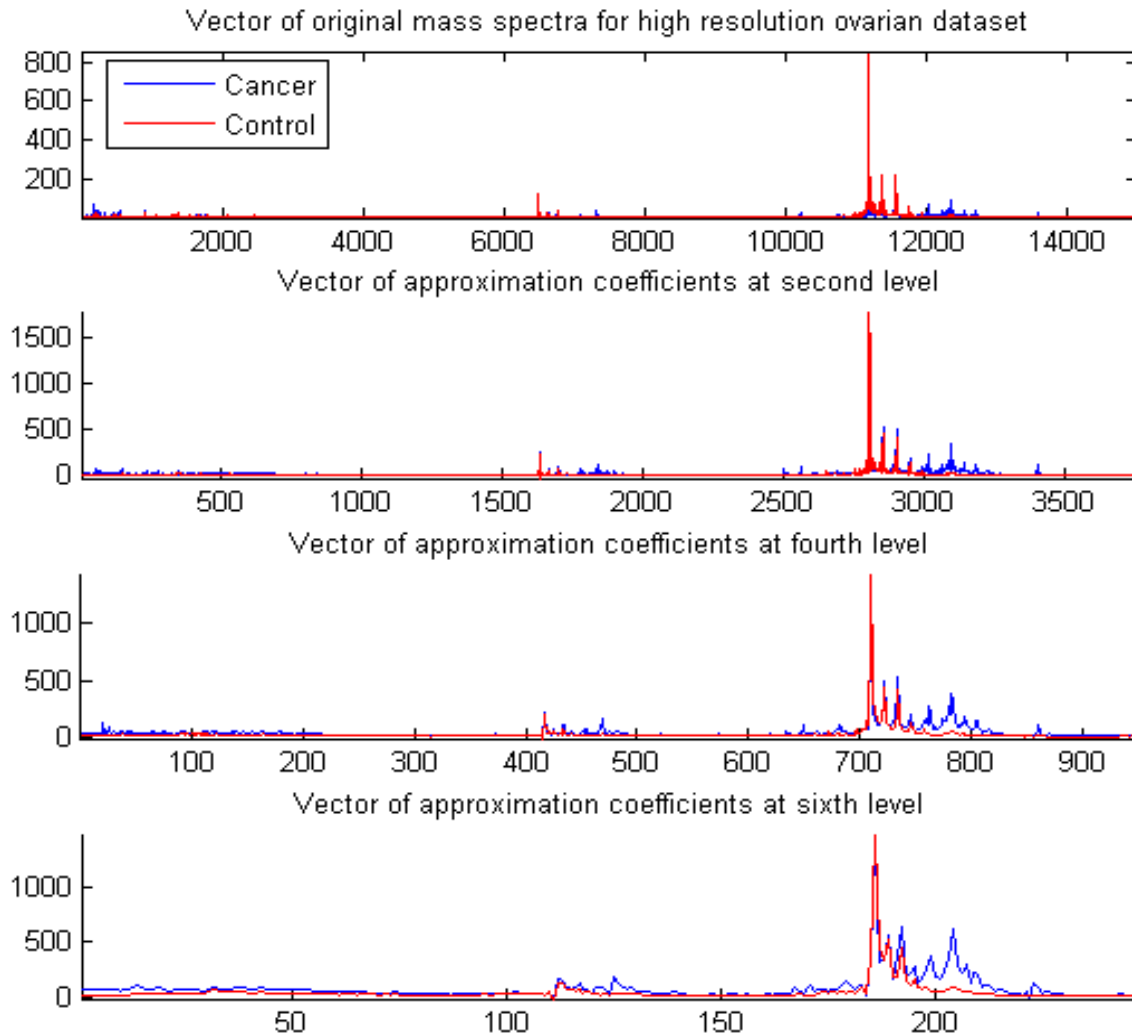
Methods	Level	Correct rate	Sensitivity	Specificity	PPV	NPV	BACC
Approximation coefficients	1	0.9989	0.9979	0.9994	0.9989	0.9988	0.9986
	2	0.9986	0.9972	0.9994	0.9989	0.9984	0.9983
	3	0.9977	0.9957	0.9988	0.9979	0.9976	0.9973
	4	0.9977	0.9965	0.9984	0.9972	0.9980	0.9974
	5	0.9986	0.9982	0.9988	0.9979	0.9990	0.9985
	6	0.9983	0.9982	0.9984	0.9972	0.9990	0.9983
Detail coefficients	4						0.9995

This Table shows the performance of low resolution ovarian dataset (8-7-02) based on three fold cross validation experiments, running fifty times (50x3CV).

Performance of different methods [2]

Methods	BACC	Sensitivity	Specificity	PPV
No FE	0.834	0.822	0.846	0.891
PCA	0.893	0.867	0.920	0.926
PCA/LDA	1.000	1.000	1.000	1.000
SFS	0.991	0.989	0.994	0.994
SBS	0.903	0.911	0.895	0.942
P-test	0.975	0.956	0.994	0.976
T-test	0.834	0.822	0.846	0.897
KS-test	0.983	0.978	0.988	0.988
NSC(20)	0.973	0.978	0.969	0.988
Boosted	0.982	0.989	0.975	0.994
Boosted FE	1.000	1.000	1.000	1.000

High resolution ovarian dataset



Best decomposition level

Energy distribution of wavelet decomposition at six levels for high resolution ovarian dataset.

Approximation coefficients	Level1		Level2		Level3		Level4		Level5		Level6	
Dimensionality	7506		3759		1886		949		481		247	
Energy distribution (percentage)	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
	99.63	0.33	99.03	0.40	95.07	8.81	62.96	5.50	58.21	6.22	48.00	8.10

Method	level	Correct rate	Sensitivity	Specificity	PPV	NPV	BACC
Approximation coefficients	1	0.9824	0.9949	0.9664	0.9742	0.9934	0.9807
	2	0.9806	0.9941	0.9633	0.9719	0.9923	0.9787
	3	0.9839	0.9965	0.9677	0.9752	0.9955	0.9821
	4	0.9800	0.9955	0.9603	0.9696	0.9940	0.9779
	5	0.9807	0.9891	0.9701	0.9768	0.9859	0.9796
	6	0.9724	0.9845	0.9569	0.9668	0.9798	0.9707
Detail coefficients [9]	3						0.9786

[This Table shows the performance for high resolution ovarian dataset based on two fold cross validation experiments, running fifty times (50x2CV).

Performance of different methods [3] of high resolution ovarian dataset.

Methods	Specificity (Mean)	Sensitivity (Mean)	BACC
VP	0.9254	0.9527	0.9390
ADAbboost	0.8959	0.9192	0.9076
1-NN	0.8320	0.8935	0.8627
3-NN	0.8131	0.8963	0.8547
ADtree	0.8180	0.8621	0.8400
J48tree	0.7785	0.8250	0.8017
Random forest	0.7760	0.8178	0.7969
Naïve Bayes	0.6591	0.7186	0.6888
SVMs	0.9330	0.9738	0.9534

Conclusions

The Comparison of properties for PCA, approximation coefficients, and detail coefficients.

Methods	Hold main components	local feature	Training samples-> feature space	Orthogonal basis	Multi-resolution	Time (253samples) 253x15154	Dimension reduction
PCA	yes	no	yes	yes	no	Out of memory	yes
WAC (Wavelet approx. coefficients)	yes	yes	no (wavelet space)	yes	yes	2.32 (seconds) six levels	yes

Reference

- [1] T.P. Conrads, M. Zhou, E.F. Petricoin III, L. Liotta, T.D. Veenstra, Cancer diagnosis using proteomic patterns, *Expert Rev. Mol. Diagn.* 3(2003) 411–420
- [2] I. Levner, Feature selection and nearest centroid classification for protein mass spectrometry, *BMC Bioinformatics* 6 (2005).
- [3] J.S. Yu, S. Ongarello, R. Fiedler, X.W. Chen, G. Toffolo, C. Cobelli, Z. Trajanoski, Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data, *Bioinformatics* 21 (2005) 2200-2209.