(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 August 2012 (16.08.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/107786 A1

(51) International Patent Classification:
G06F 19/24 (2011.01) *G06F 19/20* (2011.01)
G06K 9/62 (2006.01)

(21) International Application Number:
PCT/HR2011/000006

(22) International Filing Date:
9 February 2011 (09.02.2011)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **RUDJER BOSKOVIC INSTITUTE** [HR/HR]; Bijenicka cesta 54, 10000 Zagreb (HR).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **KOPRIVA, Ivica** [HR/HR]; Rudjer Boskovic Institute, Bijenicka cesta 54, 10000 Zagreb (HR). **JERIC, Ivanka** [HR/HR]; Rudjer Boskovic Institute, Bijenicka cesta 54, 10000 Zagreb (HR). **HADZIJA, Mirko** [HR/HR]; Rudjer Boskovic Institute, Bijenicka cesta 54, 10000 Zagreb (HR).

(74) Agent: **VUKMIR & ASSOCIATES**; Attorneys at Law, Gramaca 2L, 10000 Zagreb (HR).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) Title: SYSTEM AND METHOD FOR BLIND EXTRACTION OF FEATURES FROM MEASUREMENT DATA

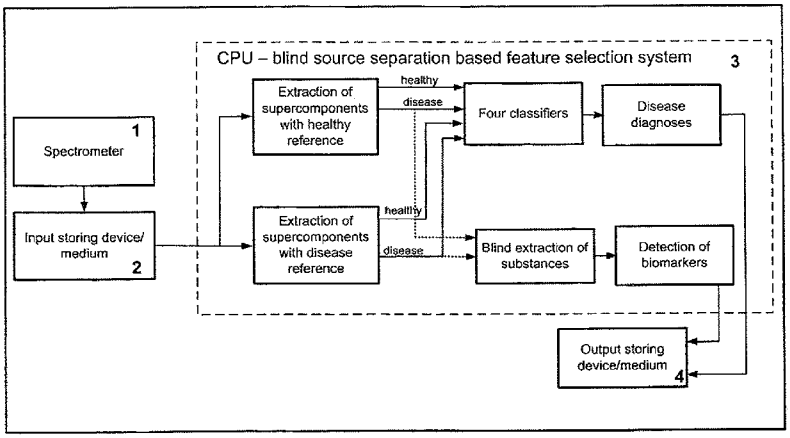

Fig. 1

(57) Abstract: A method and system for blind extraction of features from a test sample of measurement data and first and second reference samples allows to extract substances with similar weights or concentrations together as respective supercomponents. This highly simplifies blind extraction of disease-expressive substances and healthy-state-expressive substances from a sample of measurement data, and thereby effectively reduces the number of features for disease diagnosis. The invention can likewise be applied to compound activity prediction.

WO 2012/107786 A1

System and method for blind extraction of features from measurement data

TECHNICAL FIELD

The present invention relates to a system and method for blind extraction of features from a test sample of measurement data, in particular for the purpose of detecting substances that may be indicative of a disease, for biomarker detection, gene analysis, and/or compound activity prediction.

BACKGROUND OF THE INVENTION AND STATE OF THE ART

Development of methods for analysis of spectroscopic or spectrometric data acquired from biological samples that may assist in disease diagnosis and biomarker detection represents a constant challenge. In this regard, two extremes can be distinguished (R. Madsen et al., *Anal. Chim. Acta* 2010, 659:23-33): (*i*) an analysis employing pattern recognition algorithms to determine whether a suspected patient has the disease – this is known as "metabolic fingerprinting"; (*ii*) an analysis that finds chemical entities directly correlated with a certain disorder or disease that are sufficiently specific to detect the investigated disease confidently – such entities are called biomarkers, and the approach is also known as "metabolic profiling".

The sensitivity and specificity of these methods depend on a number of factors among them being: the type of biological fluid used the for analysis, the type of the spectroscopic method used for the characterization of the sample, and the type of the data analysis method employed for disease diagnosis or biomarker detection. In the last case two typical problems arise: (*i*) when pattern recognition methods, such as support vector machines or artificial neural networks, are applied for disease diagnoses spectroscopic, or spectrometric data are comprised of large number of features (even up to 30,000), compared to a much smaller number of available samples, quite often less than 100 (S. Rogers et al., *Lecture Notes in Computer Science* 2005, 3686: 183-191). Unless some type of feature selection method (also known as variable selection) is used, this leads to overfitting, i.e. causing a pattern recognition machine (classifier) to generalize (learn) on uninformative features. This decreases sensitivity

Computer Science 2005, 3686: 183-191). Unless some type of feature selection method (also known as variable selection) is used, this leads to overfitting, i.e. causing a pattern recognition machine (classifier) to generalize (learn) on uninformative features. This decreases sensitivity and specificity in disease diagnosis; (*ii*) biomarker detection from spectra of biological sample is a highly complex problem due to the fact that in some biological fluids biomarkers can be hidden among several hundreds of substances with concentrations that can vary up to few orders of magnitude (H. Mischak et al., *Mass Spectrom Rev.* 2009, 28: 703-724).

It is evident from the previous paragraph that feature selection methods are important to assist in accurate diseases diagnosis by means of classification as well as for successful detection of biomarkers. There are many feature selection/extraction methods developed, and new ones are being developing continously, see for example: *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)* Isabelle Guyon (Editor), Steve Gunn (Editor), Masoud Nikravesh (Editor), Lofti A. Zadeh (Editor), Springer, 2006. Some representative feature selection methods that constitute state of the art are briefly discussed below.

One approach to feature selection is represented by the *nearest shrunken centroids* algorithm: *"High-Dimensional problems: p >> N," Chapter 18 in: The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Second Edition (Springer Series in Statistics)* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, 2009. Here, for each feature separately the classwise mean is shrinked towards the overall mean. After the shrinking procedure, only features that have a nonzero expression for at least one class are selected for classification. This approach helps in reducing overfitting of the classification machine. However, it is purely mathematical, i.e. it lacks a deeper model-based reasoning in feature selection procedure. Thus, the efficiency of this feature selection procedure is expected to be sub-optimal.

An alternative to the *nearest shrunken centroids* type of feature selection relies on direct optimization of the vector of coefficients during a classifier training procedure. Such an approach has been described in: Z. Liu et al., *IEEE/ACM Trans. Comput. Biology and Bioinformatics* 2010, 7: 100-107. Here, the linear support vector machine classifier is designed by minimizing a loss function with a regularization term represented by the l_1-norm of the coefficient vector. This promotes sparseness of the vector of coefficients, i.e. only a

small number of features is selected during the classifier design. Said small number of features reduces the possibility of overfitting and improves the accuracy of the classification machine that is trained on a set of selected features. Additionally, when a sample is represented by the spectra the selected features represent peaks that point to interesting molecules, wherein some of them could possibly be biomarkers. From a biomarker detection standpoint, the possible drawback of this type of feature elimination is that selected features (peaks in the spectra) are not associated with any particular molecule, i.e., various combinations of selected peaks can appear in different molecules. Hence, additional knowledge is necessary to identify molecules that can possibly be biomarkers. Another disadvantage of this approach to feature selection is that classifier design and feature selection are part of the same process. Thus, this concept cannot be applied to other types of classifiers.

US Patent 7,318,051 entitled *Methods for feature selection in a learning machine* presents a method that also exploits sparseness in a feature selection process. This is done by minimizing the number of non-zero parameters of the system through l_0-norm minimization. Under certain conditions, minimization of l_p-norms, with $0 \leq p \leq 1$, is equivalent to the l_0-norm minimization method presented in US Patent 7,318,051, and the method described in the preceding paragraph belongs to the same group of sparseness-based feature selection methods. Thus, selected features are useful for learning classifiers that are robust to overfitting. However, the relation between selected features and molecules that possibly are candidates for biomarkers is not straightforward. It is hence expected that the method described in US 7,318,051 is not useful for biomarker detection.

In US Patent 7,457,048 B2 entitled *Pre-Processed Feature Ranking for a Support Vector Machine*, two types of feature extraction methods are presented: (a) a first one where the significance of particular feature is based on its influence on distance of the sample from a line that separates control from case samples; (b) a second one that looks for a sparse solution of the vector of classification coefficients. Both feature selection methods are optimized for SVM classifiers and cannot be used with other classifiers. In addition, a sparseness-based feature selection method combined with a nonlinear classifier is difficult to interpret, i.e., it is impossible to locate selected features. Feature location is important for biomarker identification because it can be associated with a specific biological function or chemical structure.

US Patent 7,676,442 B2 entitled *Selection of Features Predictive of Biological Conditions Using Protein Mass Spectrographic Data* presents a method for kernel selection for SVM classifiers invariant with respect to noise present in the data. The kernel selection process is related to the preprocessing of the mass spectral data prior to classification. This method works with SVM classifiers only. Basically the same method is also presented in US Patent Application US 2010/0205124 A1 entitled *Support Vector Machine-Based Method for Analysis of Spectral Data.* The only difference is that in US Patent 7,676,442 B2 the application target are mass spectrographic data, while in US 2010/0205124 A1 the application target are infrared spectral data.

In United States Patent Application 20100002929 entitled *Image-based methods for measuring global nuclear patterns as epigenetic markers of cell differentiation*, blind source separation in the form of nonparametric independent component analysis (ICA) is used as a learning basis of the images of the cells. This basis is then used by various methods to extract features from corresponding images of the cells. These features are subsequently used for the classification task. Feature extraction that is based on the ICA learned basis is, however, a known concept (Feature Extraction by ICA, Chapter 21 in: A. Hyvärinen, J. Karhunen, E. Oja. Independent Component Analysis, John Wiley, 2001). Thus, an ICA-based blind source separation approach to feature extraction is not based on any specific data model and is not applicable to spectral data of biological samples.

In United States Patent Application US2007176088 A1 entitled *Feature Selection in Mass Spectral Data* a method is presented for dividing mass spectra of the biological sample into feature groups for identification of biomarkers. The grouping criteria are based on information about retention time, mass isotope pattern, charge state, abundance, mass defect and the number of ions.

In United States Patent Application US2002095260 A1 entitled *Methods for efficiently mining broad data sets for biological markers* a method is presented for analysis of biological data sets for the purpose of biomarker detection. It assumes that the number of biological measurements (subjects) is at least 10 times greater than the number of observations. Thus, when the number of observations represents the number of spectral lines (that can be in the range of 10 to 20 thousands) this method requires an unrealistically large number of

measurements. For reduction of the measured data set to find possible biomarker candidates, the method relies on correlation analysis and hierarchical clustering.

In patent application WO2007145789 entitled *Method and implementation of reliable consensus feature selection in biomedical discovery*, a stand-alone method is proposed that evaluates and ranks features extracted by multiple methods for the purpose of biomarkers discovery.

Patent application WO2008037479 entitled *Feature selection on proteomic data for identifying biomarkers* presents a subset feature selection method on proteomic data, in particular high-throughput mass spectrometry data for disease diagnosis with high sensitivity and specificity as well as for biomarker discovery. In particular, the method is developed for prostate and ovarian cancer. It is a three-step process that at each step reduces the number of features and at least maintains the classification accuracy. In each step it combines existing methods for feature extraction, ranking and evaluation through use of the classifiers. The method proposed in WO2008037479 yields high accuracy (100% on ovarian cancer data set from the NCI and over 97% for prostate cancer) by linear SVM and 10-fold cross-validation. However, no double-blind cross-validation has been performed. Thus, such a high level of accuracy should be taken cautiously, since it is known that 10-fold cross-validation (as well as leave-one-out or 5-fold cross-validations) tend to yield a too optimistic classification accuracy.

In patent application WO 2007015459 entitled *Gene set for use in prediction of occurrence of lymph node metastasis of colorectal cancer*, a method is proposed and specifically developed for detection of presence or absence of lymph node metastasis of colorectal cancer.

Patent application WO2008035286 entitled *Advanced computer-aided diagnosis of lung nodules* presents a feature selection from multi-sliced computed tomography images and for detection and diagnosis of lung cancer.

United States Application US2008033899 A1 entitled *Feature selection using support vector machine classifier* presents a method for feature selection that is based on a classifier weight, whereat the feature with the smallest weight is removed from a feature set, and a support vector machine (SVM) classifier is trained again to evaluate the effect of the removed feature

on a classification performance. This feature selection scheme is, in principle, equivalent to a recursive feature elimination method. It is intimately related to the SVM classifier that is used as evaluation cost function in a features subset identification and is not applicable with other types of classifiers.

Patent application US2008025591 A1 entitled *Method and system for robust classification strategy for cancer detection from mass spectrometry data* presents a feature selection method for mass spectrometry data. The feature selection principle is based on selection of peaks of the mass spectra that are most suited to discriminate between cancer and non-cancer cases in the training set.

Patent application WO2009067655 entitled *Methods for feature selection through local learning; breast and prostate cancer prognostic markers* presents a feature selection method that employs a concept of locally linear learning that is based on maximization of margin defined locally, i.e. for each feature vector separately.

Patent application WO2005040739 entitled *System and method for spectral analysis* presents an independent component analysis (ICA) based approach to separation of spectral data into independent components. ICA cannot be applied to a two spectral data model. This is due to the fact that two spectral data model assumes that spectral data are composed from linear combination of three sets of features ("supercomponents"). Thus, two spectral data should be decomposed blindly into three supercomponents, which is a problem that cannot be solved by ICA. ICA requires that the number of spectral data available be equal or greater than number of supercomponents.

In United States Patent Application 20090287107 entitled *Analysis of eeg signals to detect hypoglycemia*, blind source separation in a form of ICA or non-zero matrix factorization is mentioned formally as a method that could possibly be used to extract features from a multiple electrode signals. This is considered as an alternative to features that are based on an average power contained in five frequency bands. Features are intended to be used for detection of hypoglycemia. The ICA-based blind source separation approach to feature extraction is not based on any specific data model and is not applicable to spectral data of biological samples.

OVERVIEW OF THE PRESENT INVENTION

In view of the problems associated with the prior art, it is the objective of the present invention to provide a method and system for blind extraction of features from a test sample of measurement data, such as spectra or gene expression profiles of biological samples or a set of molecular descriptors of compounds, that allows for a more reliable and more accurate detection of the substances in the sample.

This objective is achieved by a method for blind extraction of features from a test sample of measurement data with the features of independent claim 1, and by a system for blind extraction of features from a test sample of measurement data with the features of independent claim 14. The dependent claims relate to preferred embodiments.

A method for blind extraction of features from a test sample of measurement data according to the present invention comprises the steps of pairing said test sample with a first reference sample for said measurement data to obtain a first pairing, said first control sample pertaining to a first group of features, and pairing said test sample with a second reference sample for said measurement data to obtain a second pairing, said second control sample pertaining to a second group of features. The method further comprises the steps of decomposing said first pairing into a plurality of N sets and N corresponding weights, N being an integer no smaller than two, wherein each said set corresponds to a group of features, and wherein at least a first set corresponds to said first group of features and at least a second set corresponds to said second group of features. The method further comprises the step of decomposing said second pairing into a corresponding plurality of N sets and N corresponding weights. Again, each said set may correspond to a group of features, wherein at least a first set may correspond to said first group of features and wherein at least a second set may correspond to said second group of features.

In the terminology of the present invention, said groups of features may be called supercomponents, in accordance with the terminology introduced for describing the state of the art.

For instance, the test sample may be a mass spectrum obtained from a body fluid of a patient under investigation. The first reference sample may be a corresponding mass spectrum

acquired from a healthy subject, so that the first group of features are features associated with a healthy subject. The second reference sample may be a corresponding mass spectrum obtained from a subject having a certain disease, said second group of features thereby corresponding to features characteristic for said disease. In more generality, said first reference sample may be a control sample, and said second reference sample may be a case sample, or vice versa.

Decomposing said first pairing combining said test sample data with data from said first reference sample into said plurality of N sets and corresponding weights, and decomposing said second pairing combining said test sample data with data from said second reference sample into a corresponding plurality of N sets and N corresponding weights highly simplifies blind extraction of disease or healthy state expressive substances from a very large number of features. This is because in the decomposition according to the present invention, disease-expressive substances share similar concentrations while healthy-state-expressive substances likewise share similar concentrations. The decomposition according to the present invention allows to group substances with similar concentration profiles, thereby effectively reducing the number of features that need to be extracted for a reliable analysis of the mass spectrum acquired from the patient under investigation. As a result, the method according to the present invention helps to test for a certain disease more reliable, and with greater accuracy.

Said test sample and/or first reference sample and/or plurality of sets and/or weights may each have the form of multi-component (row or column) vectors. In practical applications, they may have a large number of components, with the number of components of the test sample and reference samples corresponding to the respective (discretized) sample sizes.

The method and system according to the present invention may be applied to a wide range of measurement data, for instance mass spectra, nuclear magnetic resonance (NMR) spectra, infrared spectra, ultraviolet (UV) spectra, Raman spectra, and electronic paramagnetic resonance spectra.

The same advantages result when the method according to the present invention is applied in the analysis of gene expression profiles to assist in disease diagnosis or for biomarker detection. In this case, the test sample and reference samples may be gene expression profiles

acquired from a patient under investigation and corresponding control (healthy) and case (disease) groups, respectively.

The data may also be sets of collections of molecular descriptors for compound activity prediction, and again the same advantages of a more reliable extraction of features result.

Preferably the number N of sets and corresponding weights is larger than 2, and in a particularly preferred embodiment N equals 3. In this case, the features extracted from said test sample may be reliably grouped into features pertaining to a healthy subject, features pertaining to a subject diagnosed with a particular disease, and neutral features that may not be easily associated with either a healthy or a disease-diagnosed subject. This particular grouping allows to focus any subsequent analysis on those features that can be clearly associated either with a disease or with a healthy state, and allows to disregard and discard those features that do not provide a clear indication in either direction. As a result, the number of features under consideration can be further reduced, thereby simplifying and speeding up the analysis.

In a preferred embodiment, said step of decomposing comprises a step of factorizing said first pairing and/or factorizing said second pairing into a plurality of N sets and N corresponding weights. Blind extraction is thereby reduced to a pair of matrix factorization problems, the first matrix factorization problem being based on a set of data obtained from said test sample and said first reference sample, and the second matrix factorization problem having the same structure, but being based on a set of data obtained from said test sample and said second reference sample. This allows to split up the task of blind extraction of features into two separate matrix factorization problems, the first problem being based on the "healthy" reference data, while the second factorization problem is based on the "disease-diagnosed" reference data. Since disease-expressive substances and healthy-state-expressive substances share similar relative concentrations, splitting up the task of feature extraction into corresponding partial factorization problems allows for a significant simplification and speed-up.

Preferably, said step of decomposing said first pairing comprises the step of representing said first pairing as a matrix factorization $(x_1, x)^T = A_1 \cdot (s_{11}, s_{12}, \ldots, s_{1N})^T$, with (x_1, x) denoting said first pairing of said test sample (x) with said first reference sample (x_1), A_1 denoting a weight

matrix, and each s_{1j} for j = 1,...,N corresponds to one of said N groups of features, and/or wherein said step of decomposing said second pairing comprises the step of representing said second pairing as a matrix factorization $(x_2, x)^T = A_2 \cdot (s_{21}, s_{22}, \ldots, s_{2N})^T$, with (x_2, x) denoting said second pairing of said test sample (x) with said second reference sample (x_2), A_2 denoting a weight matrix, and each s_{2j} for j = 1,...,N corresponds to one of said N groups of features. Here and throughout the presentation, $(.)^T$ denotes a vector transpose or matrix transpose.

The step of decomposing said first pairing preferably comprises the step of selecting said first set corresponding to said first group of features by determining the set of weights that most resembles said first reference sample, and/or selecting said second set corresponding to said second group of features by determining the set of weights that most resembles said test sample.

Correspondingly, said step of decomposing said second pairing preferably comprises the step of selecting said first set corresponding to said first group of features by determining the set of weights that most resembles said test sample, and/or selecting said second set corresponding to said second group of features by determining the set of weights that most resembles said second reference sample.

The preferred embodiment allows to reliably identify the features representative of a healthy state by comparing the corresponding weight vector obtained from the first decomposition with a healthy reference sample, while identifying the features representative of a disease by comparing the corresponding weight vector with the reference sample obtained from a disease-diagnosed subject.

The inventors have found that this provides a fast and yet reliable way of extracting those features representative of a disease, and distinguishing them from those features representative of a healthy state.

Preferably, the degree of resemblance or similarity may be evaluated in terms of an angle between a weight vector and a vector representing said first reference sample, second reference sample, or test sample, respectively. A smaller angle may correspond to a greater degree of resemblance or similarity.

The method according to the present invention preferably also comprises a step of training at least one classifier on at least four training sets gathered from said first set extracted from said first pairing, said second set extracted from said first pairing, second first set extracted from said second pairing, and said second set extracted from said second pairing. Said classifier may be applied to indicate whether a selected set of features that is present in the spectrum of a test sample relates to a disease or a healthy state. The classifier may include a pattern recognition machine or a Bayes classifier, a support vector machine classifier, a relevance vector machine classifier, a Gaussian process classifier, a classifier based on Fisher's discriminant, a boosted classifier, a naive Bayes classifier, a K-nearest neighbour classifier, or a neural network classifier.

In a preferred embodiment, the method according to the present invention further comprises the step of pairing said second set with M-1 corresponding sets obtained from M-1 distinct test samples of measurement data to obtain a third pairing, wherein M is a positive integer no smaller than two, and decomposing said third pairing into a plurality of P sets and P corresponding weights, P being an integer no smaller than 2, wherein each said set corresponds to a substance associated with one of said features.

This allows to reliably identify the substances associated with the disease-diagnosed features based on a larger set of test samples, in particular based on the second set of features diagnosed as disease-related features according to the method of the present invention as described above. Since the extraction of features according to the preferred embodiment is based on test samples from a plurality of individuals, the statistical basis is enhanced and the reliability of the identification of the substances associated with a disease is further increased.

Preferably, said step of decomposing said third pairing comprises a step of factorizing said third pairing into a plurality of P sets and P corresponding weights. This allows to reduce the determination of substances relating to the disease to a matrix factorization problem.

In a preferred embodiment, said step of decomposing said third pairing comprises a step of representing said third pairing as a matrix factorization $(u_1,u_2,\ldots, u_M)^T = V\cdot(z_1,z_2,\ldots,z_P)^T$, with $(u_1,u_2,\ldots, u_M)$ denoting said third pairing of said second set (u_1) with said M-1 corresponding

sets (u_2,..., u_M) obtained from M-1 distinct test samples, V denoting a weight matrix, and each z_j for j = 1,...,P corresponds to one of said P substances.

Preferably, said step of decomposing said first, second, and/or third pairing comprises a blind source separation, in particular an under-determined blind source separation.

In a preferred aspect, the present invention relates to a method for blind extraction of group of features, henceforth called supercomponents, from two sets of acquired spectra, wherein said blind extraction comprises the following steps:

- recording a spectrum **x** of a test sample by means of a spectrometer,
- storing said recorded spectrum **x**,
- forming two sets of two spectra $\{\mathbf{x}_1, \mathbf{x}\}$ and $\{\mathbf{x}_2, \mathbf{x}\}$, wherein $\mathbf{x}_1$ represents a reference spectrum acquired from a sample that is obtained from a healthy subject, while $\mathbf{x}_2$ represents a reference spectrum acquired from a sample that is obtained from a disease-diagnosed subject;
- representing said sets of two spectra $\{\mathbf{x}_1, \mathbf{x}\}$ and $\{\mathbf{x}_2, \mathbf{x}\}$ by a model that is defined by Equations (1) and (2):

$$\begin{pmatrix} x_1 \\ x \end{pmatrix} = A_1 \cdot \begin{pmatrix} s_{11} \\ s_{12} \\ s_{13} \end{pmatrix} = (a_{11}, a_{12}, a_{13}) \cdot \begin{pmatrix} s_{11} \\ s_{12} \\ s_{13} \end{pmatrix}, \quad (1)$$

$$\begin{pmatrix} x_2 \\ x \end{pmatrix} = A_2 \cdot \begin{pmatrix} s_{21} \\ s_{22} \\ s_{23} \end{pmatrix} = (a_{21}, a_{22}, a_{23}) \cdot \begin{pmatrix} s_{21} \\ s_{22} \\ s_{23} \end{pmatrix}, \quad (2)$$

  wherein $\{s_{11}, s_{12}, s_{13}\}$ and $\{s_{21}, s_{22}, s_{23}\}$ are row vectors that represent two sets of three supercomponents, and $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$ are column vectors that represent two sets of concentration profiles that are accompanied to related supercomponents;
- decomposing said two sets of spectra according to Equations (1) and (2) into two sets of supercomponents $\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}$ and $\{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\}$ and two sets of concentration profiles $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$ by means of an underdetermined blind source separation;

- selecting two supercomponents from each of the two sets $\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}$ and $\{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\}$ by associating them with appropriate vectors of concentration profiles $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$, wherein a disease-expressive supercomponent is extracted from the first set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by a spectrum of said test sample $\mathbf{x}$, a supercomponent that is expressive for a healthy state is extracted from the first set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by a spectrum of said reference sample $\mathbf{x}_1$, a disease-expressive supercomponent is extracted from the second set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by a spectrum of a reference sample $\mathbf{x}_2$, and a supercomponent that is expressive for a healthy state is extracted from the second set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by a spectrum of said test sample $\mathbf{x}$;
- applying at least one classifier on four training sets of disease and healthy state expressive supercomponents extracted from labelled data,
- applying said classifier to said supercomponents extracted from said test sample;
- employing supercomponents with disease expressive features that are extracted from spectra acquired from samples of disease diagnosed subjects to form a set $\{\mathbf{u}_1, \mathbf{u}_2,...,\mathbf{u}_M\}$ that is represented by a linear mixture model defined by Equation (3):

$$\begin{pmatrix} u_1 \\ \vdots \\ u_M \end{pmatrix} = V \cdot \begin{pmatrix} z_1 \\ \vdots \\ z_P \end{pmatrix} = (v_1 \cdots v_P) \cdot \begin{pmatrix} z_1 \\ \vdots \\ z_P \end{pmatrix}, \qquad (3)$$

wherein $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_P\}$ represent column vectors of concentration profiles associated with substances $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P\}$ from which disease expressive supercomponents $\{\mathbf{u}_1, \mathbf{u}_2,...,\mathbf{u}_M\}$ are composed of; and

- applying a blind source separation algorithm to $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M\}$ in Equation (3) to extract substances $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P\}$.

The latter method may be applied to the detection of disease-specific chemical compounds, such as biomarkers, which may be present in biological fluids such a urine, blood plasma, cerebrospinal fluid, saliva, amniotic fluid, bile, tears, or tissue extracts.

Said method may be used for the detection of diabetes, leukaemia, hepatitis C, Alzheimer's disease, HIV infection, coronary artery disease, depression, renal cell carcinoma, carcinoma of the urinary tract, prostate neoplasia III, ovarian cancer, prostate cancer, colon cancer, kidney cancer, Kaposi's sarcoma, benign prostatic hyperplasia, urinary tract obstruction, vacuities, diabetic nephropathy, IgA nephropathy, membranous glomerulonephritis, kidney stones, focal segmental glomerulonefroze, Fanconi's syndrome, systemic lupus erythematosus, Henoch-Schoenlein purpura, or undetected kidney disease.

Said method may also be used for the diagnosis of the state of organs during transplantation of post transplant lymphoproliferative condition, transplantation of stem cells, transplantation of hematopoietic tissue, kidney transplantation, liver transplantation, or pancreas transplantation.

According to another aspect, the present invention relates to a method for blind extraction of three groups of features, henceforth supercomponents, from two sets of two collections of gene expression profiles for disease diagnosis and biomarker detection, with the following steps:

- recording a gene expression profile **x** of a test sample,
- storing said recorded gene expression profile **x**,
- forming two sets of two gene expression profiles $\{\mathbf{x}_1, \mathbf{x}\}$ and $\{\mathbf{x}_2, \mathbf{x}\}$, wherein $\mathbf{x}_1$ represents a reference gene expression profile acquired from a sample that is obtained from a healthy subject, while $\mathbf{x}_2$ represents a reference gene expression profile acquired from a sample that is obtained from a disease-diagnosed subject;
- representing said sets of two gene expression profiles $\{\mathbf{x}_1, \mathbf{x}\}$ and $\{\mathbf{x}_2, \mathbf{x}\}$ by a model that is defined by Equations (1) and (2):

$$\begin{pmatrix} x_1 \\ x \end{pmatrix} = A_1 \cdot \begin{pmatrix} s_{11} \\ s_{12} \\ s_{13} \end{pmatrix} = (a_{11}, a_{12}, a_{13}) \cdot \begin{pmatrix} s_{11} \\ s_{12} \\ s_{13} \end{pmatrix}, \quad (1)$$

$$\begin{pmatrix} x_2 \\ x \end{pmatrix} = A_2 \cdot \begin{pmatrix} s_{21} \\ s_{22} \\ s_{23} \end{pmatrix} = (a_{21}, a_{22}, a_{23}) \cdot \begin{pmatrix} s_{21} \\ s_{22} \\ s_{23} \end{pmatrix}, \quad (2)$$

wherein $\{s_{11}, s_{12}, s_{13}\}$ and $\{s_{21}, s_{22}, s_{23}\}$ are row vectors that represent two sets of three supercomponents, and $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$ are column vectors that represent two sets of concentration profiles that are accompanied to related supercomponents;

- decomposing said two sets of spectra according to Equations (1) and (2) into two sets of supercomponents $\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}$ and $\{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\}$ and two sets of concentration profiles $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$ by means of an underdetermined blind source separation;
- selecting two supercomponents from each of the two sets $\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}$ and $\{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\}$ by associating them with appropriate vectors of concentration profiles $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$, wherein a disease-expressive supercomponent is extracted from the first set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by a spectrum of said test sample $\mathbf{x}$, a supercomponent that is expressive for a healthy state is extracted from the first set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by a spectrum of said reference sample $\mathbf{x}_1$, a disease expressive supercomponent is extracted from the second set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by a spectrum of a reference sample $\mathbf{x}_2$, and a supercomponent that is expressive for a healthy state is extracted from the second set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by a spectrum of said test sample $\mathbf{x}$;
- applying at least one classifier on four training sets of disease and healthy state expressive supercomponents extracted from labelled data,
- applying said classifier to said supercomponents extracted from said test sample;
- employing supercomponents with disease expressive features that are extracted from spectra acquired from samples of disease diagnosed subjects to form a set $\{\mathbf{u}_1, \mathbf{u}_2,...,\mathbf{u}_M\}$ that is represented by a linear mixture model defined by Equation (3):

$$\begin{pmatrix} u_1 \\ \vdots \\ u_M \end{pmatrix} = V \cdot \begin{pmatrix} z_1 \\ \vdots \\ z_P \end{pmatrix} = (v_1 \cdots v_P) \cdot \begin{pmatrix} z_1 \\ \vdots \\ z_P \end{pmatrix}, \tag{3}$$

wherein $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_P\}$ represent column vectors of concentration profiles associated with substances $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P\}$ from which disease expressive supercomponents $\{\mathbf{u}_1, \mathbf{u}_2,...,\mathbf{u}_M\}$ are composed of; and

- applying a blind source separation algorithm to $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M\}$ in Equation (3) to extract substances $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P\}$.

Preferably, said latter method may be employed in the detection of diabetes, leukaemia, hepatitis C, Alzheimer's disease, HIV infection, coronary artery disease, depression, renal cell carcinoma, carcinoma of the urinary tract, prostate neoplasia III, ovarian cancer, prostate cancer, colon cancer, kidney cancer, Kaposi's sarcoma, benign prostatic hyperplasia, urinary tract obstruction, vacuities, diabetic nephropathy, IgA nephropathy, membranous glomerulonephritis, kidney stones, focal segmental glomerulonefroze, Fanconi's syndrome, systemic lupus erythematosus, Henoch-Schoenlein purpura, or undetected kidney disease.

Said method may be employed to assist in the analysis of organs during transplantation of post transplant lymphoproliferative condition, transplantation of stem cells, transplantation of hematopoietic tissue, kidney transplantation, liver transplantation, or pancreas transplantation.

According to a further aspect, the present invention is directed at a method for blind extraction of three groups of features, henceforth supercomponents, from two sets of two collections of molecular descriptors for compound activity prediction, with the following steps:

- collecting molecular descriptors $\mathbf{x}$ of a test sample,
- storing the collected molecular descriptors $\mathbf{x}$,
- forming two sets of two collections $\{\mathbf{x}_1, \mathbf{x}\}$ and $\{\mathbf{x}_2, \mathbf{x}\}$, wherein $\mathbf{x}_1$ represents reference molecular descriptors collected from a sample that is obtained from an inactive chemical compound, while $\mathbf{x}_2$ represents reference molecular descriptors collected from a sample that is obtained from an active compound;
- representing said sets of two collections of molecular descriptors $\{\mathbf{x}_1, \mathbf{x}\}$ and $\{\mathbf{x}_2, \mathbf{x}\}$ by a model that is defined by Equations (1) and (2):

$$\begin{pmatrix} x_1 \\ x \end{pmatrix} = A_1 \cdot \begin{pmatrix} s_{11} \\ s_{12} \\ s_{13} \end{pmatrix} = (a_{11}, a_{12}, a_{13}) \cdot \begin{pmatrix} s_{11} \\ s_{12} \\ s_{13} \end{pmatrix}, \tag{1}$$

$$\begin{pmatrix} x_2 \\ x \end{pmatrix} = A_2 \cdot \begin{pmatrix} s_{21} \\ s_{22} \\ s_{23} \end{pmatrix} = (a_{21}, a_{22}, a_{23}) \cdot \begin{pmatrix} s_{21} \\ s_{22} \\ s_{23} \end{pmatrix}, \tag{2}$$

wherein $\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}$ and $\{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\}$ are row vectors that represent two sets of three supercomponents, and $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$ are column vectors that represent two sets of concentration profiles that are accompanied to related supercomponents;

- decomposing said two sets of collections of molecular descriptors according to Equations (1) and (2) into two sets of supercomponents $\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}$ and $\{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\}$ and two sets of concentration profiles $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$ by means of an underdetermined blind source separation,
- selecting two supercomponents from each of the two sets $\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}$ and $\{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\}$ by associating them with appropriate vectors of concentration profiles $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$, wherein a supercomponent expressive for an active compound is extracted from said first set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by molecular descriptors of said test sample $\mathbf{x}$, a supercomponent that is expressive for an inactive compound is extracted from said first set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by molecular descriptors of said reference sample $\mathbf{x}_1$, a supercomponent expressive for an active compound is extracted from the second set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by molecular descriptors of said reference sample $\mathbf{x}_2$, a supercomponent that is expressive for an inactive compound is extracted from the second set by associating it with the concentration profile vector that makes the smallest angle with the axis defined by molecular descriptors of said test sample $\mathbf{x}$,
- applying at least one classifier on four training sets of supercomponents expressive for active and inactive states of compounds that are extracted from labelled data,
- applying said classifier to said supercomponents extracted from said test sample;
- employing supercomponents with active state expressive features that are extracted from molecular descriptors collected from samples of active compounds to form a set $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M\}$ that is represented by a linear mixture model defined by Equation (3):

$$\begin{pmatrix} u_1 \\ \vdots \\ u_M \end{pmatrix} = V \cdot \begin{pmatrix} z_1 \\ \vdots \\ z_P \end{pmatrix} = (v_1 \cdots v_P) \cdot \begin{pmatrix} z_1 \\ \vdots \\ z_P \end{pmatrix}, \tag{3}$$

wherein $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_P\}$ represent column vectors of concentration profiles associated with the substances $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P\}$; and

- applying a blind source separation algorithm to $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M\}$ in Equation (3) to extract substances $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P\}$ from which active state expressive supercomponents $\{\mathbf{u}_1, \mathbf{u}_2,...,\mathbf{u}_M\}$ are composed.

Preferably, said method may be employed in the detection of a pattern of active state specific molecular descriptors present in a collection of molecular descriptors of a chemical compound.

Preferably, in the preceding embodiments an underdetermined blind source separation method may extract three supercomponents $\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}$ and $\{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\}$ from two pairings $\{\mathbf{x}_1, \mathbf{x}\}$ and $\{\mathbf{x}_2, \mathbf{x}\}$ by means of sparse component analysis algorithm and single component points as described in: I. Kopriva, I. Jerić, *Blind separation of analytes in nuclear magnetic resonance spectroscopy and mass spectrometry: sparseness-based robust multicomponent analysis*, Anal.l Chem., vol. 82, pp. 1911-1920, 2010, and I. Kopriva, I. Jerić, *Method of and system for blind extraction of more pure components than mixtures in 1D and 2D NMR spectroscopy and mass spectrometry combining sparse component analysis and single component points*, PCT/HR2009/000028. However, other methods developed for solving underdetermined blind source separation problems can be used for the same purpose as well.

Preferably, substances $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P\}$ in Equation (3) may be extracted from the set of disease-expressive supercomponents $\{\mathbf{u}_1, \mathbf{u}_2,...,\mathbf{u}_M\}$ or active state-expressive supercomponents, respectively by means of a blind source separation that combines sparse component analysis and single component points as described in: I. Kopriva, I. Jerić, *Blind separation of analytes in nuclear magnetic resonance spectroscopy and mass spectrometry: sparseness-based robust multicomponent analysis*, Anal. Chem., vol. 82, pp. 1911-1920, 2010.

According to a further preferred embodiment, said method may be employed to assist in disease diagnoses by means of pattern recognition algorithms to determine whether a suspected patient has the disease.

Preferably, said method may be applied to detect chemical entities directly correlated with a disease that are sufficiently specific to detect said investigated disease confidently - biomarkers.

Preferably, said method may be applied to disease diagnoses and detection of biomarkers present in biological fluids such as urine, blood plasma, cerebrospinal fluid, saliva, amniotic fluid, bile, tears, and others, or tissues or organ extracts.

The present invention likewise relates to a system for blind extraction of features from a test sample of measurement data, comprising a data input unit adapted for receiving said test sample of measurement data, a storage unit adapted to store a first reference sample for said measurement data, said first reference sample pertaining to a first group of features, and further adapted to store a second reference sample for said measurement data, said second reference sample pertaining to a second group of features, as well as a data processing unit adapted to pair said test sample with said first reference sample to obtain a first pairing, and to pair said test sample with said second reference sample to obtain a second pairing. Said data processing unit is adapted to decompose said first pairing into a plurality of N sets and N corresponding weights, N being an integer no smaller than 2, wherein each said set corresponds to a group of features, and wherein at least a first set corresponds to said first group of features and at least a second set corresponds to said second group of features. Said data processing unit is further adapted to decompose said second pairing into said plurality of N sets and N corresponding weights. Again, each said set may correspond to a group of features, wherein at least a first set may correspond to said first group of features and wherein at least a second set may correspond to said second group of features.

In a preferred embodiment, said data processing unit may be adapted to execute a method with some or all of the features as described above.

In particular, a system for blind extraction of supercomponents from two sets of two spectra according to the present invention may comprise:

- a spectrometer (1) for recording a spectrum **x** of a test sample,
- an input storing device or medium (2) for storing said spectrum **x** recorded by the spectrometer (1); and
- a processor (3), wherein said processor (3) is adapted to implement code for executing a method according to any one of the previously described embodiments based on the spectral data stored in/on the input storing device or medium (2).

The present invention likewise relates to a system for blind extraction of three groups of features, henceforth supercomponents, from two sets of two gene expression profiles, said system comprising:

- a gene chip (5) for recording a gene expression profile **x** of a test sample,
- an input storing device or medium (6) for storing said gene expression profile **x** recorded by said gene chip (5); and
- a processor (7), wherein said processor (7) is adapted to implement code for executing a method according to any one of the previously described embodiments based on the gene expression profile data stored in/on the input storing device or medium (6).

The present invention likewise relates to a system for blind extraction of three groups of features, henceforth supercomponents, from two sets of two collections of molecular descriptors for compound activity prediction, said system comprising:

- a collection (9) of molecular descriptors of a test sample **x**,
- an input storing device or medium (10) for storing molecular descriptors **x** collected by collector (9); and
- a processor (11), wherein said processor (11) is adapted to implement code for executing a method according to any of the previously described embodiments based on said collections of molecular descriptors data stored in/on the input storing device or medium (10).

Furthermore, the present invention relates to a computer-readable medium having computer-executable instructions stored thereon, which, when executed on a computer, will cause the computer to carry out a method of the present invention according to any of the preceding embodiments.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

The features and numerous advantages of a method and system according to the present invention will be best understood from a detailed description of the accompanying figures, in which:

Figure 1 schematically illustrates a block diagram of a system for blind extraction of supercomponents from spectral data and their use

for assisting in disease diagnosis and biomarker detection according to a first embodiment of the present invention;

Figures 2A and 2B — illustrate positions of vectors of concentration profiles in a plane spanned by two spectral data sets, where the second spectral data is acquired from a test sample, and the first spectral data is a reference acquired from a sample obtained from a healthy subject (Figure 2A) or a reference acquired from a sample obtained from a disease-diagnosed subject (Figure 2B);

Figure 3 — schematically illustrates blind extraction of three supercomponents (denoted symbolically by squares, rhombuses and circles) from two spectra, wherein each supercomponent is further composed of substances;

Figure 4 — schematically illustrates blind extraction of disease-expressive substances from a plurality of supercomponents extracted from spectra of samples of disease-diagnosed subjects;

Figures 5A and 5B — show a reference mass spectrum of urine of healthy mice (Figure 5A) and mass spectrum of a sample of diabetes-diagnosed mice (Figure 5B);

Figures 6A and 6B — show supercomponents extracted from two mass spectra depicted in Figures 5A and 5B, wherein Figure 6A shows a supercomponent that is expressive for a healthy state, while Figure 6B shows a supercomponent that is expressive for diabetes;

Figures 7A and 7B — show substances extracted from nine diabetes expressive supercomponents by means of a sparse component analysis that combines single component points and linear programming (Figure 7A), and l_1-norm based nonnegative matrix under-approximation (Figure 7B);

Figure 8 schematically illustrates a block diagram of a system for blind extraction of supercomponents from gene expression profile data and their use for disease diagnosis and biomarker detection according to a second embodiment of the present invention; and

Figure 9 schematically illustrates a block diagram of a system for blind extraction of supercomponents from a collection of molecular descriptors data and their use for compound activity prediction according to a third embodiment of the present invention.

A schematic block diagram of a system for blind extraction of groups of features (henceforth called "supercomponents" in accordance with the standard terminology) from two sets of two spectra by employing methods for underdetermined blind source separation according to an embodiment of the present invention is shown in Figure 1. The system consists of: a spectrometer 1 employed to acquire a spectrum from the biological sample (such as a body fluid or tissue); a storing device 2 employed to store gathered spectral data; a CPU 3 or computer where algorithms are implemented for blind extraction of supercomponents, disease diagnoses based on classification of disease-expressive and healthy-state-expressive supercomponents, blind extraction of substances from a set of disease expressive supercomponents and for biomarker detection; and an output device 4 used to store and present disease diagnoses results and biomarker candidates.

Blind extraction of three supercomponents from two sets of two spectra according to an embodiment of the present invention can algebraically be expressed as a matrix factorization problem by means of which recorded mixtures are represented by Equations (1) and (2):

$$\begin{pmatrix} x_1 \\ x \end{pmatrix} = A_1 \cdot \begin{pmatrix} s_{11} \\ s_{12} \\ s_{13} \end{pmatrix} = (a_{11}, a_{12}, a_{13}) \cdot \begin{pmatrix} s_{11} \\ s_{12} \\ s_{13} \end{pmatrix}, \tag{1}$$

$$\begin{pmatrix} x_2 \\ x \end{pmatrix} = A_2 \cdot \begin{pmatrix} s_{21} \\ s_{22} \\ s_{23} \end{pmatrix} = (a_{21}, a_{22}, a_{23}) \cdot \begin{pmatrix} s_{21} \\ s_{22} \\ s_{23} \end{pmatrix}, \tag{2}$$

where $\{s_{11}, s_{12}, s_{13}\}$ and $\{s_{21}, s_{22}, s_{23}\}$ are row vectors that represent unknown supercomponents, and $\{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, \mathbf{a}_{23}\}$ are column vectors that represent the unknown concentration profiles of related supercomponents. In both sets $\mathbf{x}$ represents a spectrum of a test sample. These may be data points indicating a relative abundance for several selected values of mass divided by charge (m/z), as acquired from a mass spectrometer, and may be written in form of a (possibly very large) vector. In the first set $\mathbf{x}_1$ represents a reference spectrum acquired from a sample that is obtained from a healthy subject, while in the second set $\mathbf{x}_2$ represents a reference spectrum acquired from a sample that is obtained from a disease-diagnosed subject. Both reference spectra can be in the same vector format as the test sample, wherein the components of a vector represents the relative abundance of a mass spectrum for various values of mass divided by charge.

Each of the three supercomponents $\{s_{11}, s_{12}, s_{13}\}$ and $\{s_{21}, s_{22}, s_{23}\}$ extracted from Equation (1) and (2) are further composed of group of components, henceforth called substances, with similar concentrations.

According to Figures 2A and 2B, which correspond to spectral data models (1) and (2), disease and healthy-state-expressive supercomponents may be identified by associating them with concentration vectors that make the smallest and the largest angles with respect to axis defined by said reference spectrum $\mathbf{x}_1$ and x_2, respectively.

Figure 2A corresponds to the data model of Equation (1), wherein second spectral data is acquired from a test sample, and first spectral data is a reference acquired from a sample obtained from a healthy subject. Figure 2B corresponds to the data model according to Equation (2), wherein second spectral data is acquired from a test sample and first spectral data is a reference acquired from a sample obtained from a disease-diagnosed subject. In the model according to Equation (1), the supercomponent that corresponds to the concentration vector closest to reference spectra is combination of features that are expressive for a healthy state, while the supercomponent that corresponds to the concentration vector closest to the test spectrum is a combination of disease-expressive features. In the model according to Equation (2), a supercomponent that corresponds to the concentration vector closest to the reference spectra is composed of disease-expressive features, while a supercomponent that corresponds to the concentration vector closest to the test spectrum is a combination of features that are expressive for a healthy state.

Hence, the interpretation of the supercomponents s_{11}, s_{12}, and s_{13} in Equation (1) is as follows. The first supercomponent s_{11} collects the disease-expressive features, while the second supercomponent s_{12} collects the healthy-state-expressive features. The third supercomponent s_{13} collects those features that cannot be reliably classified as either disease-expressive or healthy-state-expressive. In other words, Equation (1) is a decomposition of the first set of spectra $\{x_1, x\}$ into disease-expressive s_{11}, healthy-state-expressive s_{12}, and neutral s_{13} supercomponents, with corresponding weights (or concentrations) $\{a_{11}, a_{12}, a_{13}\}$.

Correspondingly, Equation (2) is a decomposition of the second set of spectra $\{x_2, x\}$ into disease-expressive s_{21}, healthy-state-expressive s_{22}, and neutral s_{23} supercomponents, with corresponding weights (or concentrations) $\{a_{21}, a_{22}, a_{23}\}$.

By means of the selection described with reference to Figures 2A and 2B, the features relating to disease-expressive and healthy-state-expressive substances can be reliably identified and extracted. Since disease-expressive substances and healthy-state-expressive substances share similar relative concentrations, substances with similar concentrations can be extracted together as one supercomponent. This highly simplifies blind extraction of disease-expressive and healthy-state-expressive substances from a very large number of features. It also makes feature extraction robust with respect to biological variability of the sample.

Blind extraction of three supercomponents from two spectra may be achieved by means of underdetermined blind source separation (uBSS). The enabling concept for the solution of such problems is known under the common name "sparse component analysis" (SCA). Theoretical foundations of the solution of the uBSS problem employing SCA are laid down in: P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representation, Signal Processing 81, 2353-2362, 2001; Y. Li, A. Cichocki, S. Amari, "Analysis of Sparse Representation and Blind Source Separation," Neural Computation 16, pp. 1193-1234, 2004; Y. Li, S. Amari, A. Cichocki, D.W.C. Ho, S. Xie, "Underdetermined Blind Source Separation Based on Sparse Representation," IEEE Trans. On Signal Processing, vol. 54, No. 2, 423-437, 2006; P. Georgiev, F. Theis, and A. Cichocki, "Sparse Component Analysis and Blind Source Separation of Underdetermined Mixtures," IEEE Trans. on Neural Networks, vol. 16, No. 4, 992-996, 2005.

For blind extraction of three supercomponents from two spectra it is possible to combine an estimation of the concentration matrix on a set of single component points by data clustering with an estimation of the supercomponents by convex optimization, as described in: I. Kopriva, I. Jerić, *Blind separation of analytes in nuclear magnetic resonance spectroscopy and mass spectrometry: sparseness-based robust multicomponent analysis*, Anal. Chem., vol. 82, pp. 1911-1920, 2010. This exemplary method for underdetermined blind source separation will now briefly be outlined. With reference to Equations (1) and (2), the matrix of mixing vectors or concentration vectors is estimated from a set of features (mass/charge ratios, genes, molecular descriptors, etc.) where only one supercomponent is dominant. These components may be detected by means of a geometric criterion that is based on the notion that the real and imaginary part of the mixture samples point either in the same or opposite directions at the features where only one component is dominant. Any standard clustering procedure (k-means clustering, c-means clustering, fuzzy c-means clustering, spectral clustering, hierarchical clustering, etc.) can be employed to cluster the set of features with single component dominance. Concentration vectors ($\mathbf{a}_{11}$, $\mathbf{a}_{12}$, $\mathbf{a}_{13}$) and/or ($\mathbf{a}_{21}$, $\mathbf{a}_{22}$, $\mathbf{a}_{23}$) are represented by the cluster centers (centroids). Since in Equation (1) and (2) the number of concentration vectors correspond to the number of supercomponents and equals 3, the number of clusters is known in advance and equals 3. Once the matrix of concentration vectors has been estimated, supercomponents $\{\mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}\}$ in Equation (1) and $\{\mathbf{s}_{21}, \mathbf{s}_{22}, \mathbf{s}_{23}\}$ in Equation (2) are obtained by solving a linear system of two equations in three unknowns at each feature. This system is solvable if at each feature at least one supercomponent has zero value, i.e. the vector comprised of the entries of three supercomponents at the particular feature should be sparse.

Many methods exist for solving an underdetermined system of linear equations under sparseness assumption, see for example: Tropp, J.A., Wright, S.J., 2010, "Computational Methods for Sparse Solution of Linear Inverse Problems", Proc. of the IEEE 98, 948-958. One representative method that may be employed in the context of the present invention is the ℓ_1-regularized least square problem that at the feature index i reads as $\hat{\mathbf{s}}_{(i)} = \arg\min_{\mathbf{s}_{(i)}} \frac{1}{2}\left\|\hat{\mathbf{A}}\mathbf{s}_{(i)} - \mathbf{x}_{(i)}\right\|_2^2 + \lambda\left\|\mathbf{s}_{(i)}\right\|_1$, where λ represents a regularization constant enforcing a sparse solution. The interior-point method presented in Kim S.J. *et al.* (2007), "An interior-point method for large-scale ℓ_1-regularized least squares", *IEEE J. Sel. Topics Signal Proc.*, 1, 606-617, may be used to solve this ℓ_1-regularized least square problem.

Once the four supercomponents relating to disease-expressive and healthy-state-expressive features have been identified and extracted as described above, they may be used for assisting in disease diagnosis by applying a previously trained pattern recognition algorithm or classifier to them. The supecomponents relating to the neutral features may be discarded at this stage, but they may be stored for later analysis in variances of the preferred embodiment described herein.

Examples for pattern recognition algorithms are Bayes classifier, a support vector machine (SVM), a relevance vector machine (RVM), a Gaussian process classifier, a classifier based on Fisher's discriminant, a boosted classifier, a naive Bayes classifier, a K-nearest neighbour classifier, a neural network classifier, etc. A diagnosis method that is robust with respect to biological variability of the sample is obtained by selecting as the output of the four classifiers the one with the highest accuracy achieved in the cross-validation phase.

The set of supercomponents composed of the disease-expressive features may be decomposed further into less complex combinations of features that are referred to herein as substances and may be used for identification of disease-specific biomarkers.

This may be achieved by collecting disease-expressive supercomponents $\{u_1, u_2, \ldots, u_M\}$ acquired from M different samples of disease-diagnosed subjects and representing them by a linear mixture model defined by Equation (3):

$$\begin{pmatrix} u_1 \\ \vdots \\ u_M \end{pmatrix} = V \cdot \begin{pmatrix} z_1 \\ \vdots \\ z_P \end{pmatrix} = (v_1 \cdots v_P) \cdot \begin{pmatrix} z_1 \\ \vdots \\ z_P \end{pmatrix}, \qquad (3)$$

wherein $V = (v_1, \ldots, v_P)$ represents a matrix of column vectors of concentration profiles associated with the substances $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_P\}$. Thus, applying a blind source separation algorithm to $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M\}$ in Equation (3) enables to extract substances $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_P\}$. These substances are potential candidates for disease-specific biomarkers and may subsequently be subjected to a biomarker identification procedure.

Figure 3 schematically illustrates blind extraction of three supercomponents (denoted symbolically by squares, rhombuses, and circles, respectively) from two spectra. Each supercomponent is further composed of substances that are expressive for disease (squares),

healthy state (circles), or neutral (rhombuses). Different shading and texture within each supercomponent symbolically denote different substances. While the upper mixture corresponds to the first set of spectra $\{x_1, x\}$, the lower mixture corresponds to the second set of spectra $\{x_2, x\}$. Blind extraction of supercomponents according to Equations (1) and (2) yields a decomposition into supercomponents of disease-expressive, healthy-state-expressive, and neutral substances.

Figure 4 schematically illustrates the blind extraction of disease-expressive substances from a plurality of supercomponents extracted from spectra of samples of disease-diagnosed subjects according to Equation (3). Different shading and texture symbolically denote different substances. As can be taken from Figure 4, blind extraction according to Equation (3) yields a decomposition of the supercomponents associated with disease-diagnosed samples into less complex combinations of individual substances, which can then be used further for identification of disease-specific biomarkers.

Figures 5A (upper figure) and 5B (lower figure) respectively show experimental mass spectra of urine samples of healthy mice and diabetes-diagnosed mice. According to Equation (1), the mass spectrum shown in Figure 5A can be used as a healthy reference, while the mass spectrum shown in Figure 5B represents the spectrum of a test sample. According to the model of Equation (2), the mass spectrum shown in Figure 5B can be used as a disease-reference while the mass spectrum shown in Figure 5A then represents the spectrum of a test sample.

Figures 6A (upper figure) and 6B (lower figure) show two supercomponents extracted from the two mass spectra shown in Figures 5A and 5B in accordance with the data model of Equation (1), where the mass spectrum shown in Figure 5A serves as a healthy reference and the mass spectrum shown in Figure 5B serves as a test. According to the interpretation of the data model of Equation (1), the supercomponent that contains features which are expressive for a healthy state as shown in Figure 6A corresponds to the concentration vector which is closest to the reference spectrum, as explained above with reference to Figure 2A. The disease-expressive supercomponent is shown in Figure 6B and corresponds to the concentration vector which is closest to the spectrum of a test sample, as likewise explained with reference to Figure 2A.

Figures 7A and 7B show substances extracted from nine diabetes-expressive supercomponents. The substances shown in Figure 7A are extracted by means of a sparse component analysis that combines single component points and linear programming, while the substances shown in Figure 7B are extracted by means of ℓ_1-norm based non-negative matrix underapproximation, as explained above.

The invention has been described above with reference to the blind extraction of supercomponents from two sets of spectra, such as mass spectra. However, the invention is by no means limited to this specific example, and may be employed whenever the blind extraction of features from a test sample of measurement data is desired. For instance, the method according to the present invention may likewise be employed for the blind extraction of supercomponents from collections of gene expression profiles for disease diagnosis and biomarker detection, or for blind extraction of supercomponents from sets of collections of molecular descriptors for compound activity prediction. Various further applications will become apparent to those skilled in the art. The method according to the present invention equally applies to all such applications. Only the physical interpretation of the reference sample and test samples, the groups of features of supercomponents and the features itself may be different. For instance, a supercomponent relating to healthy features in the analysis of spectra or gene expression profiles may correspond to an inactive state of a chemical compound, whereas a disease-diagnosed supercomponent may correspond to an active state of a chemical compound.

A schematic block diagram of a system for blind extraction of supercomponents from two sets of two gene expression profiles that is defined by Equations (1) and (2) and employing methods for underdetermined blind source separation according to an embodiment of the present invention is shown in Figure 8. The system consists of: a gene chip 5 used to acquire gene expression profiles from a biological sample; a storing device 6 used to store gathered gene expression profiles data; a CPU 7 or computer where algorithms are implemented for blind extraction of supercomponents, disease diagnoses based on classification of disease-expressive and healthy-state-expressive supercomponents, a blind extraction of substances from a set of disease-expressive supercomponents and for biomarker detection; and an output device 8 used to store and present disease diagnoses results and biomarker candidates.

A schematic block diagram of a corresponding system for blind extraction of supercomponents from two sets of two collections of molecular descriptors of chemical compounds that is defined by Equations (1) and (2) and employing methods for underdetermined blind source separation according to an embodiment of the present invention is shown in Figure 9. The system consists of: a collector of molecular descriptors 9 used to acquire molecular descriptors data from the samples of chemical compounds; a storing device 10 used to store gathered molecular descriptors data; a CPU 11 or computer where algorithms are implemented for: a blind extraction of supercomponents and compound activity prediction based on classification of active and inactive state-expressive supercomponents, blind extraction of substances from a set of active state-expressive supercomponents and for detection of pattern of molecular descriptors that is specific for the active state; and an output device 12 used to store and present activity prediction results and candidates for active state specific pattern of molecular descriptors.

The feature selection method proposed in the patent application herein has been successfully tested on the ovarian cancer mass spectra of serum samples. The data were downloaded from: http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp, and have been published in: E.F. Petricoin et al., "Use of proteomic patterns in serum to identify ovarian cancer," The Lancet, **359**, 572-577. The data set contains 100 control samples and 100 cancer samples. To extract supercomponents one control and one cancer sample were used as reference ones. All the classifiers were applied to standardized data having zero mean and unit variance. When a linear support vector machine classifier trained on supercomponents expressive for control and extracted from a set according to Equation (1) with a control reference sample is used, a sensitivity of 93.4% ±3.4% and a specificity of 88.2% ±5.2% have been achieved in two-fold cross-validation evaluated over 100 random splittings. Using linear SVM classifier trained on supercomponents expressive for control and extracted from a set according to Equation (2) with a cancer reference sample yields a sensitivity of 92.5%±3.7% and a specificity of 91.2%±5.0 %.

A special embodiment of the method for blind extraction of three groups of features from two sets of two spectra has been also tested on the prostate cancer mass spectra of the serum samples. The data were downloaded from: http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp, and have been published in: E. F.

Petricoin III et al., "Serum Proteomic Patterns for Detection of Prostate Cancer", Journal of the National Cancer Institute (2002) 94, 1576-1578. The data set contains 63 control samples with prostate-specific-antigen (PSA) less than 1, 26 cancer samples with PSA between 4 and 10 and 43 cancer samples with PSA greater than 10. To extract supercomponents one control and one cancer sample were used as reference ones. All the classifiers were applied to standardized data with zero mean and unit variance. When a linear SVM classifier trained on supercomponents expressive for cancer state and extracted from a set according to Equation (1) with a control reference sample is used, a sensitivity of 98.9±2.1% and a specificity of 98.6±2.5% have been achieved in two-fold cross-validation. Using a nonlinear SVM classifier with a polynomial kernel of degre 2 yields a sensitivity of 98.6%±2.6% and a specificity of 98.6%±2.6 %.

A special embodiment of the method for blind extraction of three groups of features from two sets of two spectra has been further tested on colon cancer gene expression profiles data of the tissue samples. The data were downloaded from: http://genomics-pubs.princeton.edu/oncology/affydata/index.html, and have been published in: U. Alon et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proc. Natl. Acad. Sci. USA, vol. 96, pp. 6745-6750, 1999. In this case, the data set contains 22 control samples and 40 cancer samples, whereas each sample contains 2000 gene expression levels. To extract supercomponents one control and one cancer sample were used as reference ones. All the classifiers were applied to standardized data with zero mean and unit variance. When a nonlinear SVM classifier with a Gaussian kernel with variance σ^2=1000 and trained on supercomponents expressive for healthy state and extracted from a set according to Equation (2) with a cancer reference sample is used, a sensitivity of 91.1±4.9% and a specificity of 81.0±9.9% has been achieved in two-fold cross-validation. Using a nonlinear SVM classifier with polynomial kernel of degre 2 yields a sensitivity of 90.7%±6.6% and a specificity of 76.4%±12.3 %. Using a KNN classifier with K=7 yields a sensitivity of 92.2% ±4.9% and a specificity of 77.6%±9.9% has been achieved in two-fold cross-validation. A greater specificity can be achieved if data set is more balanced, i.e. having number of control samples approximately equal to the number of case samples.

A special embodiment of the method for blind extraction of three groups of features from two sets of two spectra has been tested furthermore on the diabetes mass spectra data of the urine samples of the NOD mices. The data were prepared in the in house laboratories at the Ruđer Bošković Institute and were comprised of 10 control and 10 diabetic mices. Mass spectra were acquired by a HPLC-MS triple quadruple instrument equipped with an autosampler (Agilent Technologies, USA) operating in a positive ion mode. To extract supercomponents one control and one diabetes sample were used as references ones. Thus, 9 control and 9 diabetes samples remained for cross-validation. Due to the very small sample size leave-one-out cross-validation has been performed in this case. A nonlinear SVM classifier with a RBF kernel achieved 100% sensitivity and 100% specificity.

The detailed description and the accompanying figures merely serve to illustrate the invention and the associated technical effects, but should not be understood to limit the invention in any sense. The scope of the invention is to be determined solely on the basis of the accompanying claims.

Claims

1. A method for blind extraction of features from a test sample (x) of measurement data, said method comprising the steps of:
pairing said test sample (x) with a first reference sample (x_1) for said measurement data to obtain a first pairing (x_1,x), said first reference sample (x_1) pertaining to a first group of features;
pairing said test sample (x) with a second reference sample (x_2) for said measurement data to obtain a second pairing (x_2,x), said second reference sample (x_2) pertaining to a second group of features;
decomposing said first pairing (x_1,x) into a plurality of N sets (s_{11},s_{12},...,s_{1N}) and N corresponding weights (a_{11},a_{12},...,a_{1N}), N being an integer no smaller than two, wherein each said set (s_{1j}) corresponds to a group of features, and wherein at least a first set (s_{11}) corresponds to said first group of features and at least a second set (s_{12}) corresponds to said second group of features; and
decomposing said second pairing (x_2,x) into a corresponding plurality of N sets (s_{21},s_{22},...,s_{2N}) and N corresponding weights (a_{21},a_{22},...,a_{2N}).

2. The method according to claim 1, wherein said measurement data is a mass spectrum acquired from a mass spectrometer, or a gene expression profile, or a collection of molecular descriptors for compound activity description.

3. The method according to claim 1 or 2, wherein said step of decomposing comprises a step of factorizing said first pairing (x_1,x) and/or factorizing said second pairing (x_2,x) into a plurality of N sets and N corresponding weights.

4. The method according to any of the preceding claims, wherein said step of decomposing said first pairing comprises the step of representing said first pairing as a matrix factorization $(x_1, x)^T = A_1 \cdot (s_{11},s_{12},...,s_{1N})^T$, with ($x_1$, x) denoting said first pairing of said test sample (x) with said first reference sample (x_1), A_1 denoting a weight matrix, and each s_{1j} for j = 1,...,N corresponds to one of said N groups of features, and/or wherein said step of decomposing said second pairing comprises the step of representing said second pairing as a matrix factorization $(x_2, x)^T = A_2 \cdot (s_{21},s_{22},...,s_{2N})^T$, with ($x_2$, x)

denoting said second pairing of said test sample (x) with said second reference sample (x_2), A_2 denoting a weight matrix, and each s_{2j} for j = 1,...,N corresponds to one of said N groups of features.

5. The method according to any of the preceding claims, wherein said step of decomposing said first pairing (x_1, x) comprises the step of selecting said first set (s_{11}) corresponding to said first group of features by determining the set of weights that most resembles said first reference sample (x_1), and selecting said second set (s_{12}) corresponding to said second group of features by determining the set of weights that most resembles said test sample (x).

6. The method according to any of the preceding claims, wherein said step of decomposing said second pairing (x_2, x) comprises the step of selecting said first set (s_{21}) corresponding to said first group of features by determining the set of weights that most resembles said test sample (x), and selecting said second set (s_{22}) corresponding to said second group of features by determining the set of weights that most resembles said second reference sample (x_2).

7. The method according to claim 5 or 6, wherein a degree of resemblance is evaluated in terms of an angle between a weight vector and a vector representing said first reference sample (x_1), second reference sample (x_2), or test sample (x), respectively, with a smaller angle corresponding to a greater degree of resemblance.

8. The method according to any of the preceding claims, further comprising a step of training at least one classifier on at least four training sets gathered from said first set (s_{11}) extracted from said first pairing (x_1, x), said second set (s_{12}) extracted from said first pairing (x_1, x), said first set (s_{21}) extracted from said second pairing (x_2, x), and said second set (s_{22}) extracted from said second pairing (x_2, x).

9. The method according to any of the preceding claims, further comprising the step of pairing said second set with M-1 corresponding sets obtained from M-1 distinct test samples of measurement data to obtain a third pairing, wherein M is a positive integer no smaller than two, and decomposing said third pairing into a plurality of P sets and P

corresponding weights, P being an integer no smaller than two, wherein each said set corresponds to a substance associated with one of said features.

10. The method according to claim 9, wherein said step of decomposing said third pairing comprises a step of factorizing said third pairing into a plurality of P sets and P corresponding weights.

11. The method according to claim 9 or 10, wherein said step of decomposing said third pairing comprises a step of representing said third pairing as a matrix factorization $(u_1,u_2,\ldots, u_M)^T = V\cdot(z_1,z_2,\ldots,z_P)^T$, with $(u_1,u_2,\ldots, u_M)$ denoting said third pairing of said second set (u_1) with said M-1 corresponding sets $(u_2,\ldots, u_M)$ obtained from M-1 distinct test samples, V denoting a weight matrix, and each z_j for $j = 1,\ldots,P$ corresponds to one of said P substances.

12. The method according to any of the preceding claims, wherein said step of decomposing said first pairing (x_1,x) and/or second pairing (x_2,x) and/or third pairing pairing $(u_1,u_2,\ldots, u_M)$ comprises a blind source separation, in particular an under-determined blind source separation.

13. The method according to any of the preceding claims, wherein said features are features that may be indicative of ovarian cancer, prostate cancer, colon cancer, or diabetes.

14. A system for blind extraction of features from a test sample (x) of measurement data, comprising:
a data input unit (2,6,10) adapted for receiving said test sample (x) of measurement data;
a storage unit (2,6,10) adapted to store a first reference sample (x_1) for said measurement data, said first reference sample (x_1) pertaining to a first group of features, and further adapted to store a second reference sample (x_2) for said measurement data, said second reference sample (x_2) pertaining to a second group of features; and
a data processing unit (3,7,11) adapted to pair said test sample (x) with said first reference sample (x_1) to obtain a first pairing (x_1,x), and to pair said test sample (x) with said second reference sample (x_2) to obtain a second pairing (x_2,x);
said data processing unit (3,7,11) further adapted to decompose said first pairing (x_1,x) into a plurality of N sets $(s_{11},s_{12},\ldots,s_{1N})$ and N corresponding weights $(a_{21},a_{22},\ldots,a_{2N})$, N

being an integer no smaller than two, wherein each said set (s_{1j}) corresponds to a group of features, and wherein at least a first set (s_{11}) corresponds to said first group of features and at least a second set (s_{12}) corresponds to said second group of features; and
said data processing unit (3,7,11) further adapted to decompose said second pairing (x_2,x) into a corresponding plurality of N sets (s_{21},s_{22},...,s_{2N}) and N corresponding weights (a_{11},a_{12},...,a_{1N}).

15. The system according to claim 14, wherein said data processing unit is further adapted to execute a method according to any of the claims 1 to 13.

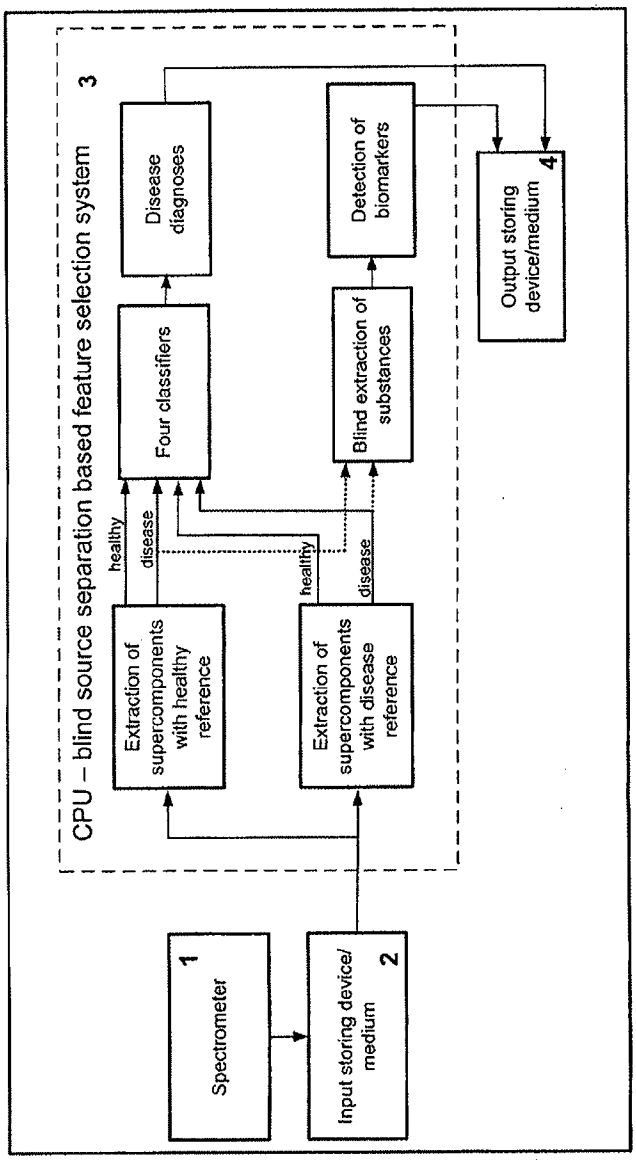
CPU – blind source separation based feature selection system
3
Spectrometer
1
Input storing device/ medium
2
Extraction of supercomponents with healthy reference
Extraction of supercomponents with disease reference
healthy
disease
healthy
disease
Four classifiers
Blind extraction of substances
Disease diagnoses
Detection of biomarkers
Output storing device/medium
4

Fig. 1

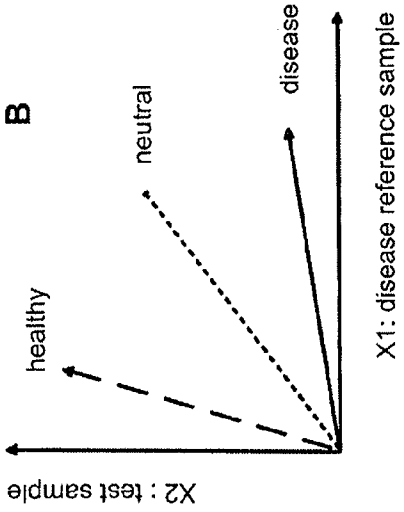
B
healthy
neutral
disease
X2 : test sample
X1: disease reference sample

Fig. 2B

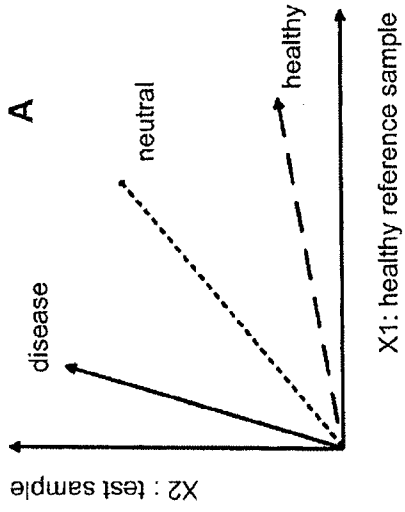
A
disease
neutral
healthy
X2 : test sample
X1: healthy reference sample

Fig. 2A

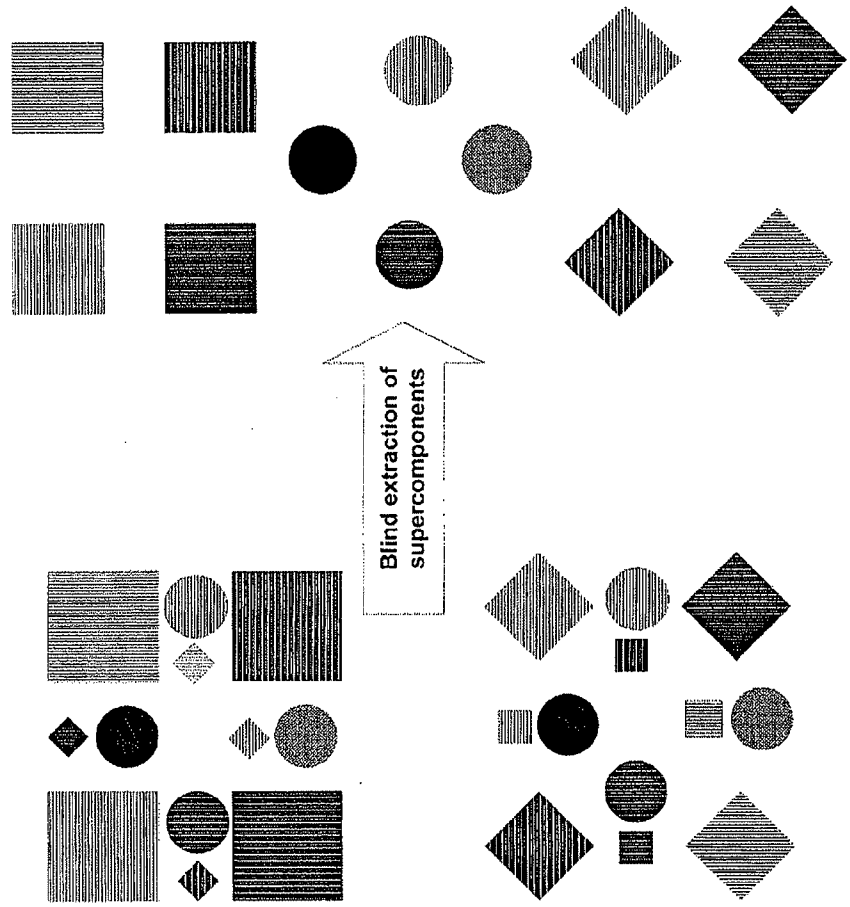

Fig. 3

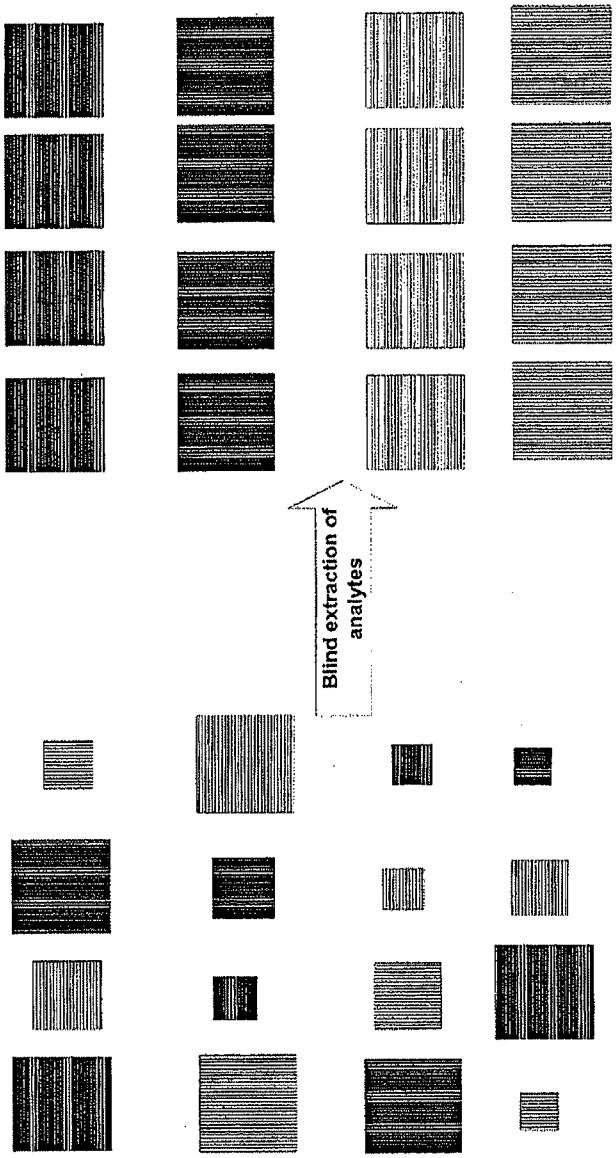

Fig. 4

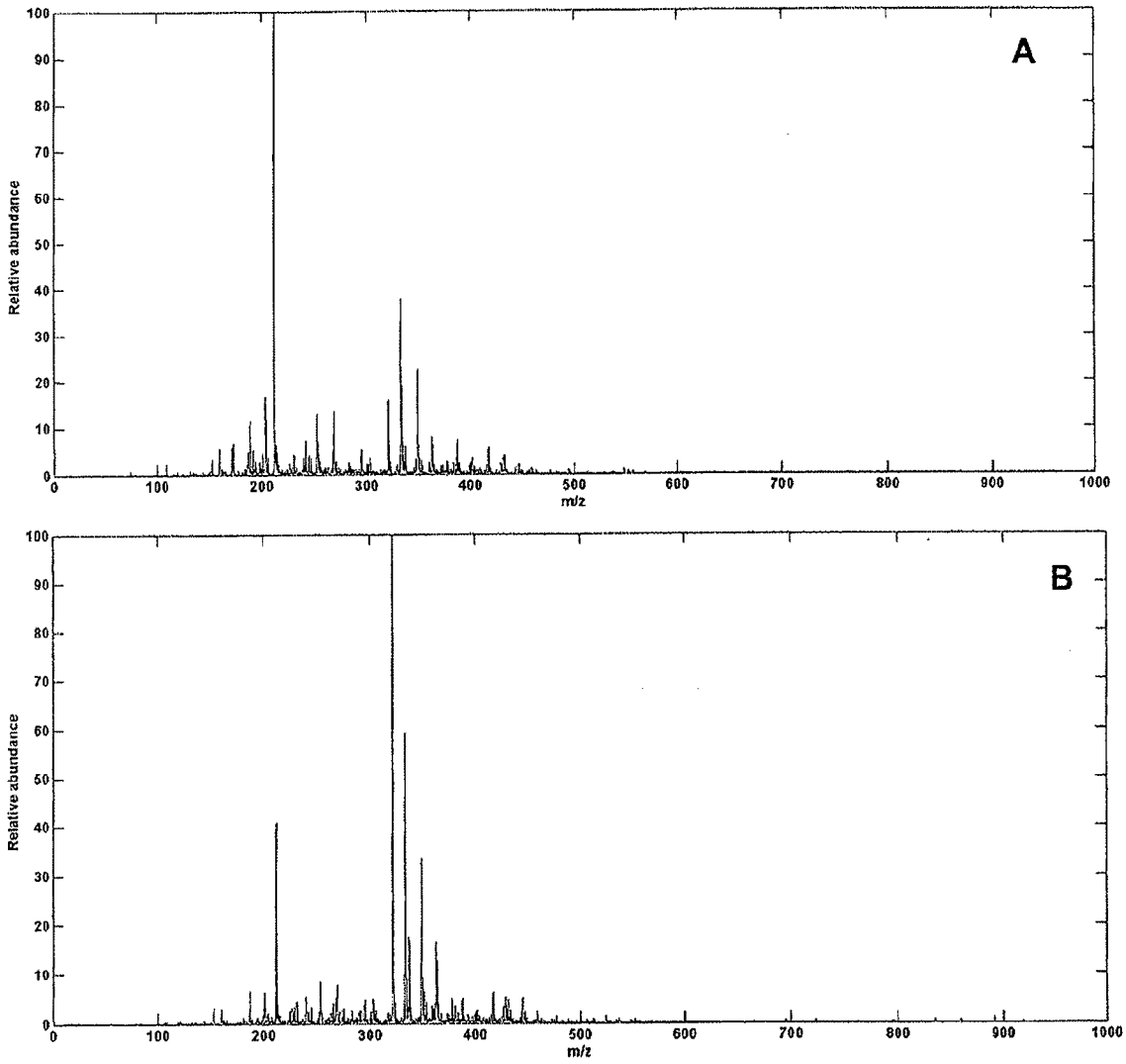
A
B
Relative abundance
m/z
0
10
20
30
40
50
60
70
80
90
100
0
100
200
300
400
500
600
700
800
900
1000

Fig. 5

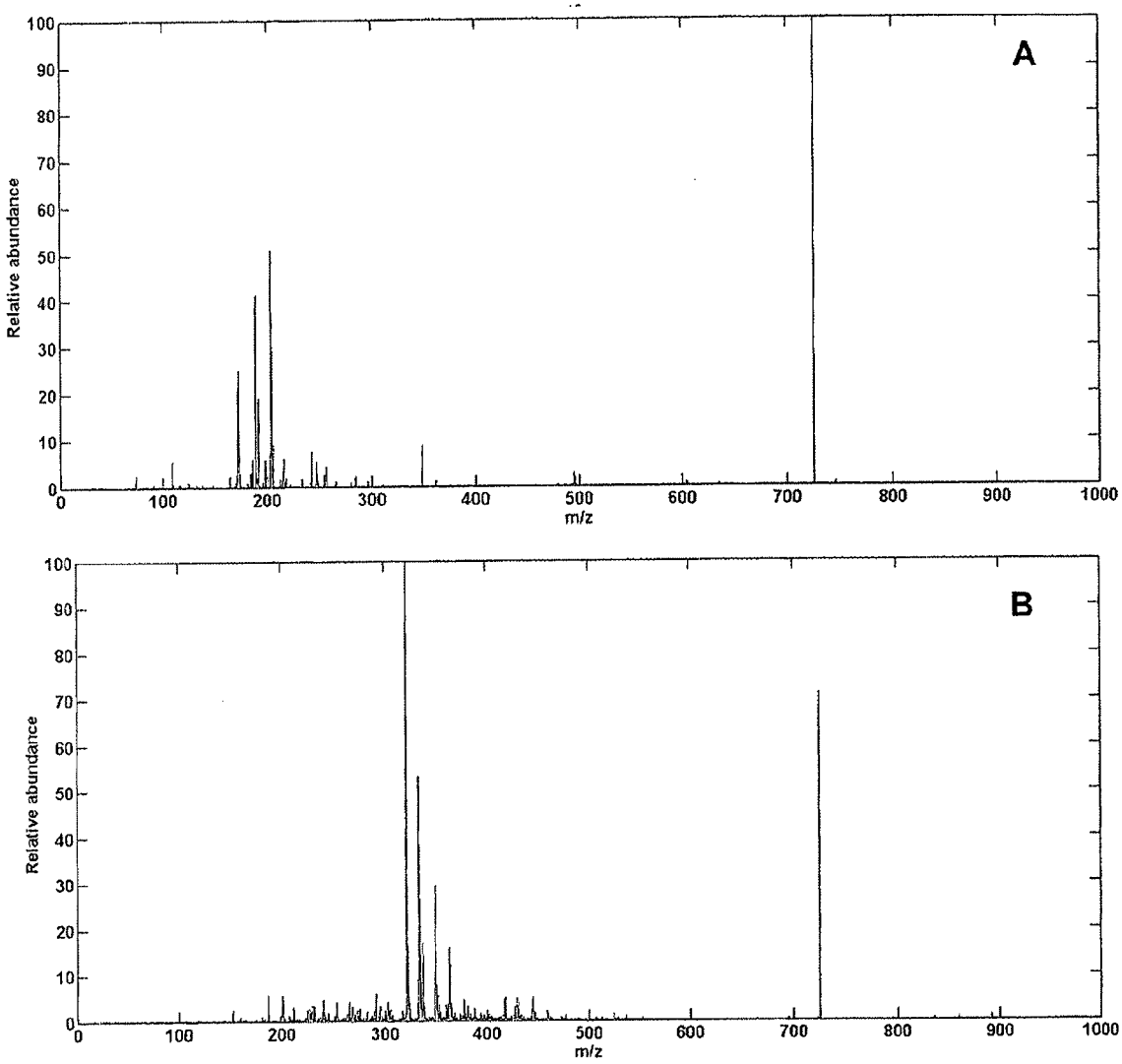
A
B
Relative abundance
m/z
0
10
20
30
40
50
60
70
80
90
100
200
300
400
500
600
700
800
900
1000

Fig. 6

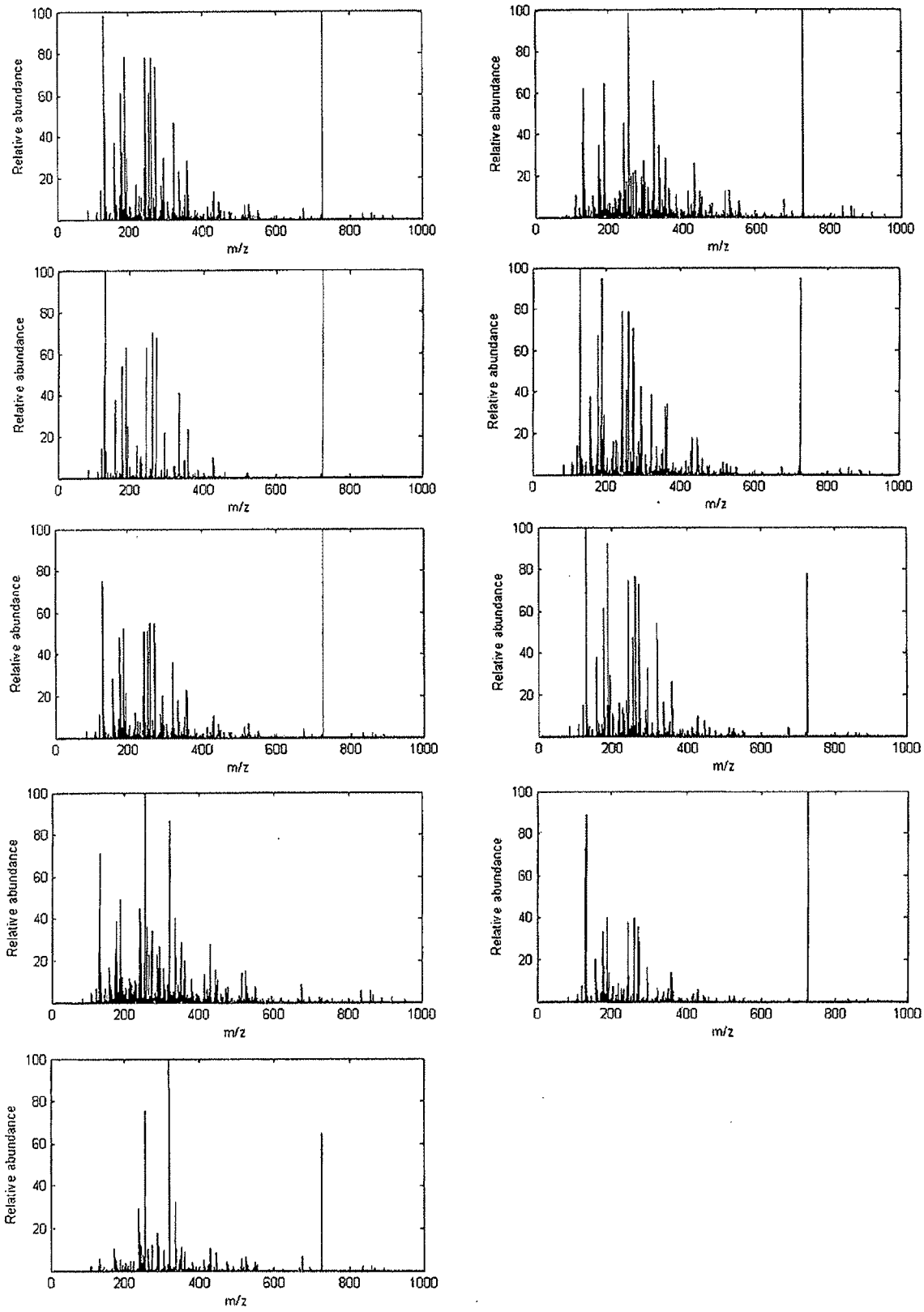

Fig. 7A

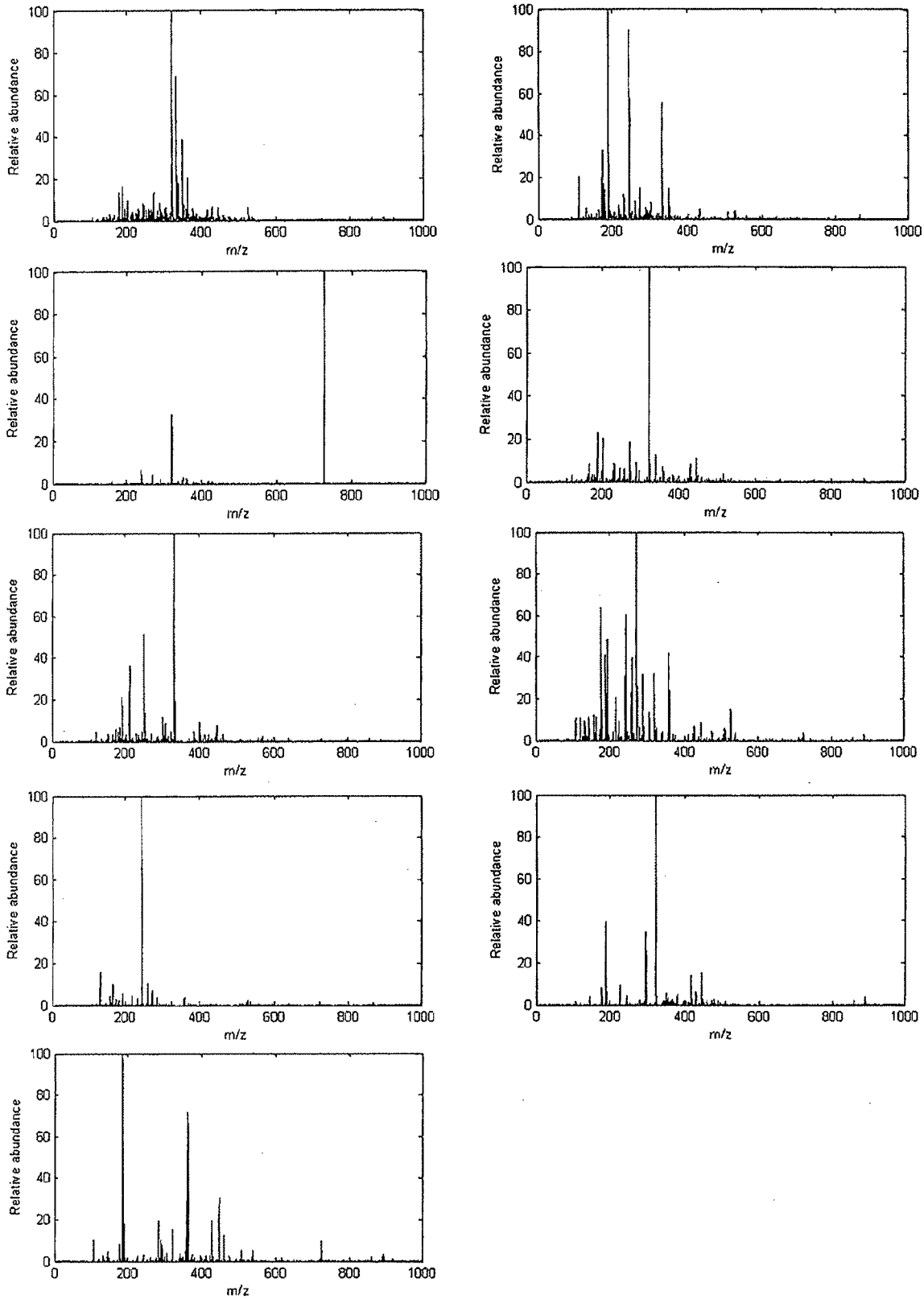

Fig. 7B

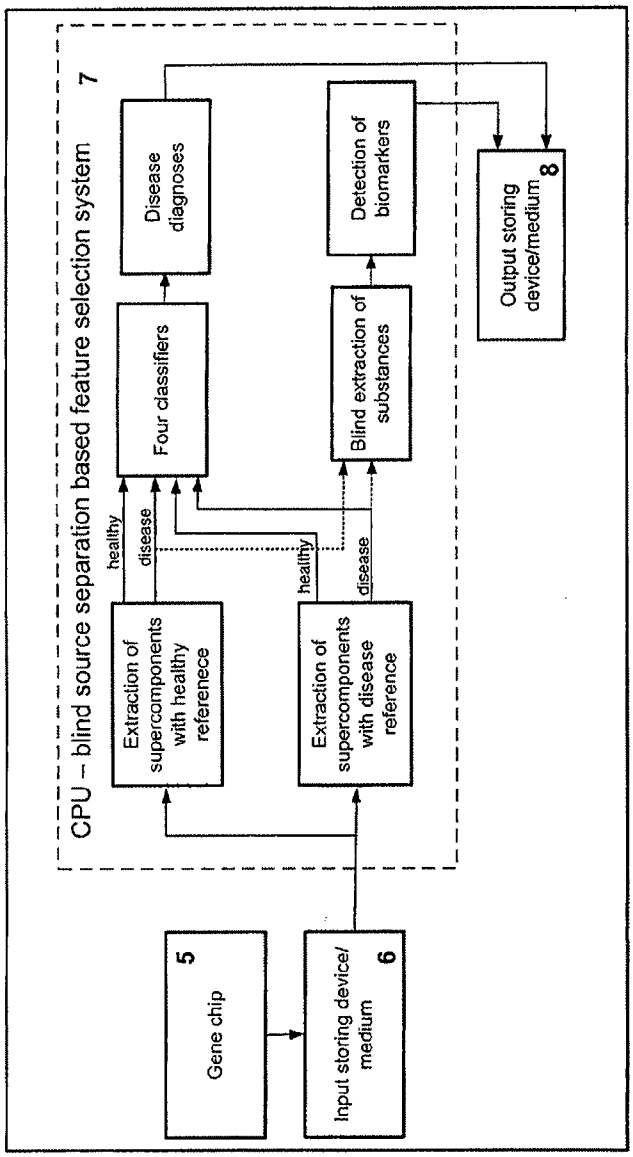
CPU – blind source separation based feature selection system
7
Gene chip
5
Input storing device/ medium
6
Extraction of supercomponents with healthy refernece
Extraction of supercomponents with disease reference
healthy
disease
healthy
disease
Four classifiers
Blind extraction of substances
Disease diagnoses
Detection of biomarkers
Output storing device/medium
8

Fig. 8

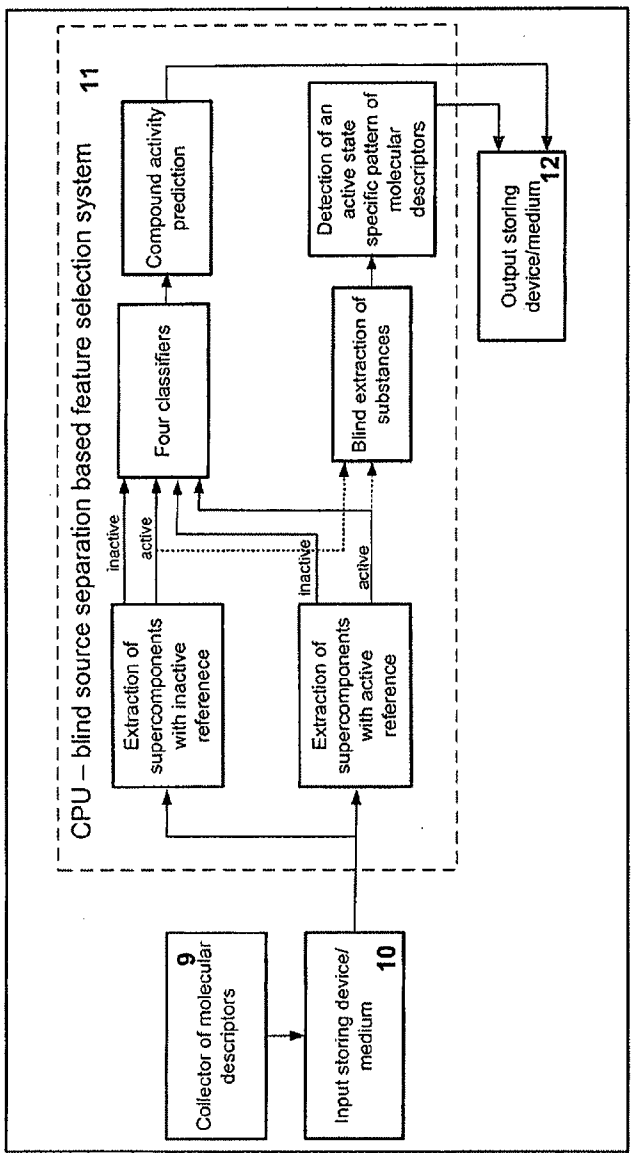
CPU – blind source separation based feature selection system 11
Collector of molecular descriptors 9
Input storing device/ medium 10
Extraction of supercomponents with inactive referenece
Extraction of supercomponents with active reference
inactive
active
inactive
active
Four classifiers
Blind extraction of substances
Compound activity prediction
Detection of an active state specific pattern of molecular descriptors
Output storing device/medium 12

Fig. 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/HR2011/000006

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. [X] Claims Nos.: 1-13
because they relate to subject matter not required to be searched by this Authority, namely:

Rule 39.1(i) PCT - Mathematical method

2. [ ] Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. [ ] Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. [ ] As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. [ ] As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. [ ] As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. [ ] No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

[ ] The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

[ ] The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

[ ] No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No
PCT/HR2011/000006

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/24 G06K9/62
ADD. G06F19/20

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2003/216868 A1 (HO MING-HSIU [US]) 20 November 2003 (2003-11-20) abstract; claims 1-32; figures 1-3 paragraphs [0005] - [0019] paragraphs [0116] - [0154] -----	14,15
X	BYUNG-SOO KIM ET AL: "Prostate cancer classification processor using DNA computing technique", IEICE ELECTRONICS EXPRESS IEICE JAPAN, vol. 6, no. 10, 2009, pages 581-586, XP002660799, ISSN: 1349-2543 page 582, last paragraph - page 583, paragraph 1 page 585, paragraph 3 figure 2 -----	14,15

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
7 October 2011	18/10/2011

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Swarén, Peter

INTERNATIONAL SEARCH REPORT

International application No
PCT/HR2011/000006

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2003/225526 A1 (GOLUB TODD R [US] ET AL) 4 December 2003 (2003-12-04) abstract; claims 1-20; figure 5 paragraphs [0015] - [0020] paragraphs [0038], [0039] paragraphs [0050] - [0061] -----	14,15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/HR2011/000006

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2003216868 A1	20-11-2003	AU 7730900 A	30-04-2001
		WO 0123614 A1	05-04-2001
		US 6505125 B1	07-01-2003
US 2003225526 A1	04-12-2003	WO 03041562 A2	22-05-2003