

# Ellipsoid clustering machine: a front line to aid in disease diagnosis

DOI: 10.3395/reciis.v1i2.Sup.101en



*Paulo Costa  
Carvalho*

Programa de engenharia  
de sistemas e computação,  
COPPE, Universidade Federa-  
l do Rio de Janeiro, Rio  
de Janeiro, Brazil  
carvalhopc@cos.ufrj.br



*Juliana de  
Saldanha da  
Gama Fischer*

Instituto de química da  
Universidade Federal do Rio  
de Janeiro e Rede Proteomi-  
ca do Rio de Janeiro, Rio de  
Janeiro, Brazil  
juli\_f@iq.ufrj.br

*Valmir C. Barbosa*

Programa de engenharia de sistemas e computação,  
COPPE, Universidade Federal do Rio de Janeiro, Rio  
de Janeiro, Brazil  
valmir@cos.ufrj.br

*Wim Degrave*

Laboratório de Genômica Funcional e Bioinformáti-  
ca, Instituto Oswaldo Cruz, Rio de Janeiro, Brazil  
wdegrave@fiocruz.br

*Maria da Gloria da Costa  
Carvalho*

Instituto de Biofísica Carlos Chagas Filho, Universi-  
dade Federal do Rio de Janeiro, Rio de Janeiro, Brazil  
mgccosta@biof.ufrj.br

*Gilberto Barbosa Domont*

Instituto de química da Universidade Federal do Rio  
de Janeiro e Rede Proteomica do Rio de Janeiro, Rio  
de Janeiro, Brazil  
gilberto@iq.ufrj.br

## Abstract

This study presents a new machine learning strategy to address the disease diagnosis classification problem that comprises an unknown number of disease classes. This is exemplified by a software called Ellipsoid Clustering Machine (ECM) that identifies conserved regions in mass spectrometry proteomic profiles obtained from control subjects and uses these to estimate classification boundaries based on sample variance. The software can also be used for visual inspection of data reproducibility. ECM was evaluated using mass spectrometry protein profiles obtained from serum of Hodgkin's disease patients (HD) and control subjects. According to the leave-one-out cross validation, ECM completely separated both groups based only on the information derived from four selected mass spectral peaks. Classification details and a 3D graphical model showing the separation between the control subject cluster and HD patients is also presented. The software is available on the project website together with online interactive models of the dataset and an animation demonstrating the method.

## Keywords

Mass spectrometry, machine learning, pattern recognition, clustering, Hodgkin's disease, proteomics

## Introduction

### Biomarkers and proteomics

During the last 40 years the possibility of early cancer detection employing biomarkers came to the forefront as a promising field to transform medical diagnosis, methods for measuring disease progression and response to treatment (CONRADS et al., 2004). However, the pursuit for a single biomarker to discriminate a unique pathology has been unsuccessful until today. Even PSA (prostate specific antigen), used to diagnose prostate cancer when highly expressed in men, can sometimes give misleading results, besides not being specific (TROYER et al., 2004).

A goal of proteomics is to distinguish between various states of a biological system to identify protein expression differences. Mass spectrometry based proteomics has been employed for differential protein expression analysis through various techniques including stable isotope labeling, ion current values, and numbers of tandem mass spectra (spectral “counting”) to name a few (LIU et al., 2004). Among the first and widely adopted proteomic techniques for biomarker discovery stands the 2-dimensional gel electrophoresis (2DE). Traditionally 2-DE studies were limited to comparative analysis but two important advances lead this technique into the “omics” era: the use of immobilized pH gradients (GORG et al., 1998) to create more reproducible gel patterns and the use of mass spectrometry to identify differentially expressed proteins (BJELLQVIST et al., 1982; HENZEL et al., 1993; WESTERMEIER et al., 1983). 2-DE contrasts different collections of proteins, (e.g. from cells or tissues), by comparing their migration according to their molecular weight and isoelectric point. Differences between the resulting protein spot collections are determined by using gel pattern analysis software. Recent developments in this field add fluorescent tags to label proteins from different samples and separate them in the same gel with a technique called differential gel electrophoresis (DIGE) (Unlu et al., 1997; VON et al., 2001). This dramatically improved the comparison of gel patterns and minimized reproducibility issues. Differentially expressed proteins can be excised from the gel and the tryptic peptide mass fingerprint (PMF) analyzed by mass spectrometry. The mass spectra data from the PMFs are used to identify the proteins which could give clues for patient treatment and can be used as biomarkers for early diagnosis (WEINGARTEN et al., 2005). Certainly, the greatest drawback of the 2DE is that it is extremely laborious and difficult to automate.

In 2002, mass spectra generated by surface enhanced laser desorption ionization time of flight (SELDI-TOF) coupled to computer algorithms, identified a set of key proteins that, according to the authors, could discriminate control subjects (CS) from ovarian cancer patients (ARDEKANI et al., 2002). The advantage of this approach over the 2DE strategy is that gels are no longer required, opening the way to high throughput studies. Later, a similar approach was used for breast cancer, using

SELDI technology with “unified maximum separability analysis” (LI et al., 2002). Another work discriminated prostate cancer from CS using decision trees with boosting techniques (QU et al., 2002) and classical statistical methods. SELDI involves the analysis of small sets of proteins, pre-selected by their affinity properties with the SELDI plate, however, depletion of proteins can result in loss of potential biomarkers and changes in the proteomic pattern (MEHTA et al., 2003).

The pursuit for the identification of multiple biomarkers to assist in the early diagnosis and prognosis of disease, and the construction of probabilistic models is crucial, especially to assist in the indication for the start of a treatment and to aid in decisions for the clinical regimen for improved chances of success. However, the identification of *bona fide* sets of biomarkers challenges the field of proteomics, requiring more sensitivity and quantification capacity than existing techniques (gel electrophoresis – 1D and 2D; chromatography online with tandem mass spectrometry). This also challenges the science of artificial intelligence for pattern recognition. Several factors limit such developments, such as: availability of small numbers of clinical samples with a gold standard diagnosis, high cost of equipment and reagents, the high number of parameters per sample, considerable variability between samples in the same class, limitations in the reproducibility of proteomic techniques for the detection and simultaneous quantification of thousands of proteins, and the lack of knowledge of a probability density function describing the variables representing the expression level of each protein for the case under study. The understanding of the interplay between the different components of a biological system and the patterns of differential qualitative and quantitative expression of such is still far away, even with all the current “omics” efforts.

### The pattern recognition problem

The construction of mathematical models to allow a machine to learn from experience and make inferences has long been a topic of discussion and philosophical debate. The supervised learning problem comprises the construction of machines that can learn, with a specialist, and then successfully classify future events. This approach is described as follows (VAPNIK, 1995):

A certain phenomenon generates events  $x$  randomly and independently according to a probability density function  $p(x)$ . These events are classified as belonging to one of the  $k$  existing classes according to a specialist. For the sake of simplicity, let  $k = 2$ ; however this assumption can be generalized to higher values of  $k$ , since by subdivision each can be divided into two classes as well. The specialist performs classification according to a conditional probability density function  $p(y|x)$  where  $y = \{+1, -1\}$  ( $y = +1$  indicates that the specialist labeled event  $x$  as belonging to the positive class, and  $y = -1$  for the negative class  $(-1)$ ). The properties of the event generating phenomenon and the specialist’s decision rules are unknown, however both exist.

Let  $C$  be the set of functional dependencies  $[F(x)]$

that serve as decision rules for the classification problem at hand. The functions are represented in their parametric form as  $F(x, \mathbf{a})$  where  $\mathbf{a}$  is a parameter belonging to the set  $\varphi$ . The value  $\mathbf{a}^*$  specifies the function  $F(x, \mathbf{a}^*)$ . The set  $\varphi$  is arbitrary, and can be composed of scalars, vectors or abstract elements. All the functions of  $C$  are indicating functions (*i.e.* having output values limited to +1 or -1). After observing  $l$  pairs

$$x_1, y_1, \dots, x_l, y_l$$

(the event is represented by  $x$  and the classification according to the instructor by  $y$ ), one should choose, among the class of indicating functions  $F(x, \mathbf{a})$ , the function having its probability of classification with minimal difference when compared with the specialist.

In other words, the minimum of the functional

$$\alpha) \sum_k^l \int (y_k - F(x, \alpha))^2 p(y_k|x)p(x),$$

should be obtained to minimize the risk. Given these facts, the classification problem is reduced to minimize the expected risk in the light of the empirical data.

## Pattern recognition and biomarkers

One of the most challenging tasks in classification through machine learning is to find a method applicable to large scale multi-class problems where the features (in this case biomolecules) and classes (here represented by the control group and the “pathology group”) are huge. Various studies in the literature, similar to the ones above, are mostly **dichotomic**, always discriminating between control subjects and patients with a given pathology. These approaches usually employ a supervised learning algorithm to train over a dataset or a selected array of up or down-regulated features (disease associated biomolecules) and establish a separating decision boundary between two classes.

The long-term goal of these studies is to develop specialist systems capable of diagnosing whether a biological sample originated from a diseased subject or not. The most successful approach in practice is to convert the multiclass classification problem to various binary classification problems and proceed with “one against all” or “all pairs” classification strategies (MAO et al., 2005; NIIJIMA et al., 2005; XU et al., 2007). However, if the described specialist system is faced with an unknown class (a disease that the specialist system was never trained to recognize), the existing set of binary classifiers may fail to satisfactorily detect the pathology and can output a Type II error (classify erroneously a patient as healthy). Given the existence of innumerable pathologies and that false negative classification is the “worst mistake” a classifier can perform for the nature of the problem at hand, the development of heuristics for this problem remains open.

## Methods and algorithm

This work introduces a new rationale to serve as a front line for classifying biological sample profiles, given the multiclass nature of the disease diagnosis dilemma. Our method aims at the identification of protein sets whose expression remains conserved in control subjects, having the expectation that some of these proteins could be altered in patients. Differently than mapping both up and down-regulated biomolecules, this procedure delineates a “pathology free” domain (conserved domain) in a feature space and could serve as a simple and straight forward first step for disease diagnosis.

The proposed specialist system would firstly use a classifier based on “conserved domains” to evaluate the probability for an unknown sample to belong to the “healthy” class or not. If the sample lies outside the conserved region boundaries, only then, the specialist system would rely on its collection of specialized (binary) classifiers to further try and classify the pathology at hand. The traditional approach that directly applies the various binary classifiers could lead to a false conclusion when facing a new class since none of the classifiers would be trained to recognize the new disease. On the other hand, our approach could have a higher chance of inferring if an unknown pathology is possibly present since it was trained over conserved regions of protein profiles from control subjects; therefore, it could still be able to alert for the possible presence of a new disease classes.

Here we developed the classification algorithm, and present a proof of principle of the concepts above. First, we acquired mass spectra profiles from serum of control subjects and patients with Hodgkin’s disease. We recall that Hodgkin’s disease (HD) belongs to a group of cancers called lymphomas that may occur in a single lymph node, a group of lymph nodes or in other parts of the lymphatic system such as bone marrow and spleen. HD tends to spread in an orderly way from one group of lymph nodes to the next.

We then developed the algorithm named ellipsoid clustering machine (ECM) having roots in the previously described concepts to search for the conserved regions in the mass spectral protein profiles of the control subjects. Finally, the algorithm used the leave-one-out (LOO) cross validation to evaluate whether it could correctly classify among control subjects and patients with Hodgkin’s disease. Mass spectral peaks that could correspond to putative HD biomarkers were also tracked using the new feature selection method described above.

The dataset used for this work originated from 30 samples obtained from healthy blood donors and 30 samples from HD patients that were collected at the Clementino Fraga Filho Federal University Hospital at Rio de Janeiro. Diagnosis and histological classification were confirmed by a hematopathologist, according to WHO criteria. Patient evaluation included a complete history, physical examination, complete blood differential count, biochemical profile, HIV serology, chest radiography, computerized tomography of the chest, abdomen and bone marrow biopsy. The sera were stored as aliquots

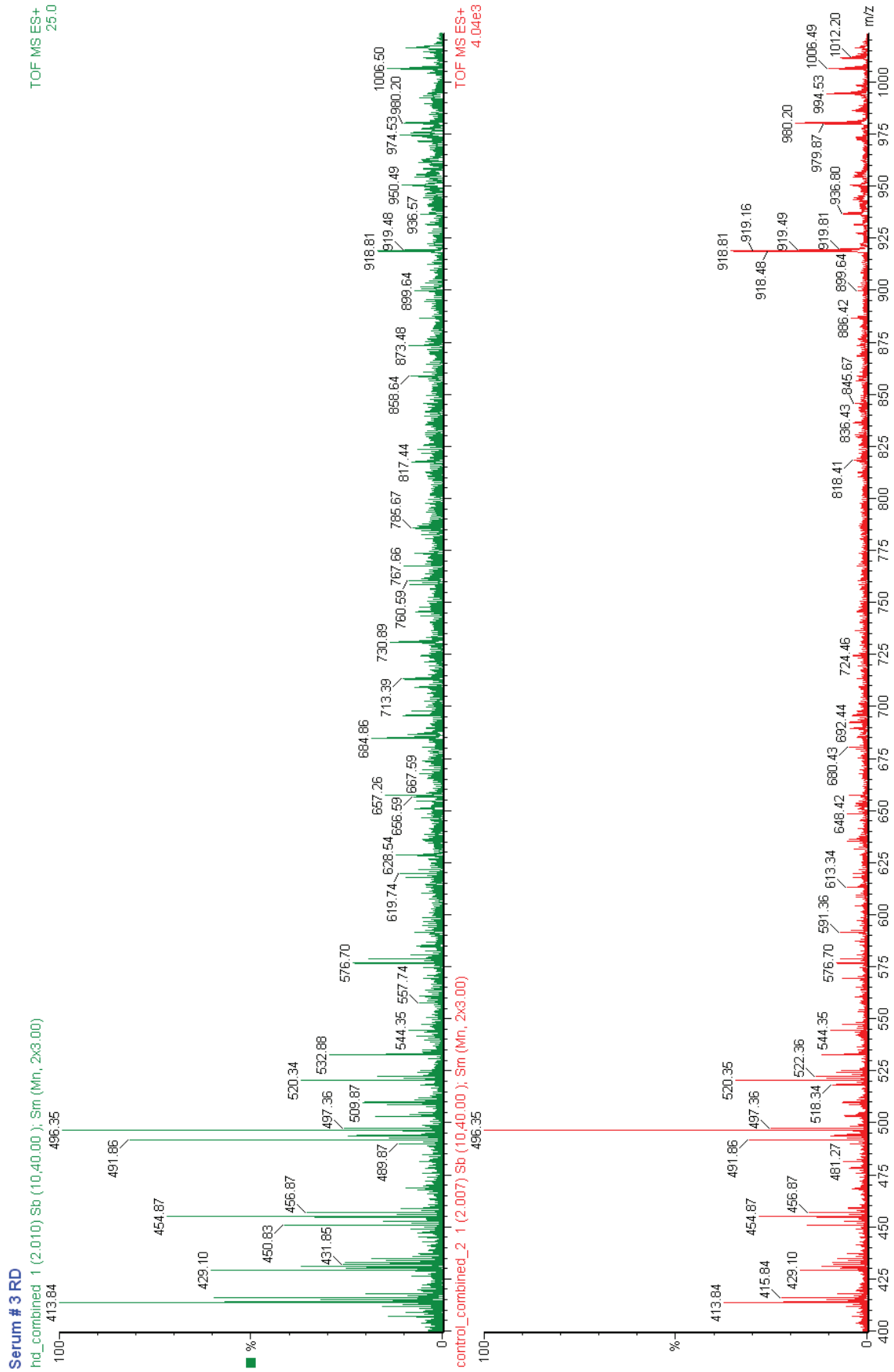


Figure 1 - The red and green mass spectra are examples of proteomic profiles obtained from the serum of a Hodgkin's disease patient and a control subject respectively. The y axis represents ion intensity and the x axis represents the mass to charge ratio of the detected ions.



at  $-80^{\circ}\text{C}$ . Demographic, tumor staging, and pathological information about the patients were then stored in a database. Sera from case patients were obtained at or after diagnosis but before treatment. All CS were free of cancer on the basis of clinical history and physical examination and no additional imaging approaches or routine marker assays were performed. All participants provided written consent for this study and the project was approved by the Federal University of Rio de Janeiro Institutional Ethical Review Board. A 1D gel analysis was then carried out to search for evidence of differentially expressed proteins and to serve as a pre-screening for all samples (CARVALHO et al., 2005). The procedure for acquiring mass spectra profiles used in this work was described previously (CARVALHO et al., 2007). The dataset analyzed also originated from the cited work. It is important to note that two mass spectra were acquired for every subject and then averaged to diminish MS noise; the mass spectra data were then binned to 1 Dalton windows by summing intermediate values. Examples of proteomic profiles can be seen in Figure 1.

## Results

The ECM algorithm was applied to the proteomic profile dataset to define the “conserved regions” based on the control subject class according to a univariate approach. This is performed as follows: for each mass spectral peak bin, the intensity of the corresponding mass spectral peak from each CS is used to map the corresponding data point to a one dimensional feature space. After all the CS's are mapped, decision boundaries are extended from each CS's data point until a predetermined number of CS's are engulfed by these boundaries. Since this process is carried out in a one dimensional space, the decision boundaries will be composed of “lines”. A list of the mass spectral profile regions that are most conserved can then be pointed as the ones whose decision boundaries had to be extended the least to satisfy such criteria. Further details of the algorithm and its source code are made available in the project website.

After selecting the conserved regions of the mass spectral profiles, the final hyper ellipsoid classification boundaries can be modeled. This is achieved by using only the data from regions of the mass spectra that were marked as conserved. From then on, ECM maps the CS to the feature space according to the mass spectral peak intensities. We note that now, this feature space has the same cardinality as the number of spectrum regions selected and the intensity of each spectral peak bin is used to map the CS according to a coordinate value. Hyper-ellipsoids originating from each CS are then positioned and extended, similar to the process above, having axis growth rates proportional to the variance of each respective dimension in a loop process. The growth of all ellipsoids ceases when every ellipsoids' center is engulfed by a user determined number of ellipsoids originating from other CS. Classification is performed by checking whether data from a new spectrum is positioned within, or out of the hyper-ellipsoid boundaries.



Figure 2 - Ellipsoids are extended from CS represented in Euclidian feature space and define a “Hodgkins disease free domain”. The red spheres represent data from HD patients; all of them are located outside the cancer free domain.

According to the leave-one-out cross validation for the dataset from our Hodgkin's disease patients, ECM correctly classified all CS and all HD patients. Figure 2 shows the decision boundary created based on the CS data represented by the blue ellipsoid cluster; HD patients are represented as small red spheres. We observe that the ellipsoid axe sizes are proportional to data variance for each respective direction. The radii of the red spheres are equal to an arbitrary constant used for mere illustration. A result worthy of note is that, in general, distant red spheres represented patients in an advanced, disseminated stage of the Hodgkin's disease while spheres closer to the ellipsoid cluster represented patients in a localized early stage of HD.

To visually evaluate how well the selected conserved regions can discriminate CS against a determined pathology, a 3D viewer is also made available. The viewer is capable, when working with 3 dimensions (3 MS peaks bins), of displaying ellipsoids representing CS in blue and patients in red spheres. The center positioning of each ellipsoid in the feature space is given by the normalized mass spectra intensity of each respective biomarker for a given subject. The internet browser should be used to view the VRML model (Virtual Reality Modeling Language); however a VRML viewer must be previously installed. The Cortona VRML client is suggested since it is freely available for download at <http://www.parallelgraphics.com/products/cortona/>. 3D interactive models are available on the project website.

## Discussion

### Pattern recognition and bioinformatics

Two main issues characterize feature selection challenges in the domain of bioinformatics: the large input

dimensionality, and limitations in the dataset size. To deal with these problems, various feature selection techniques have been designed by experts from the machine learning and data mining fields. To date, the overall

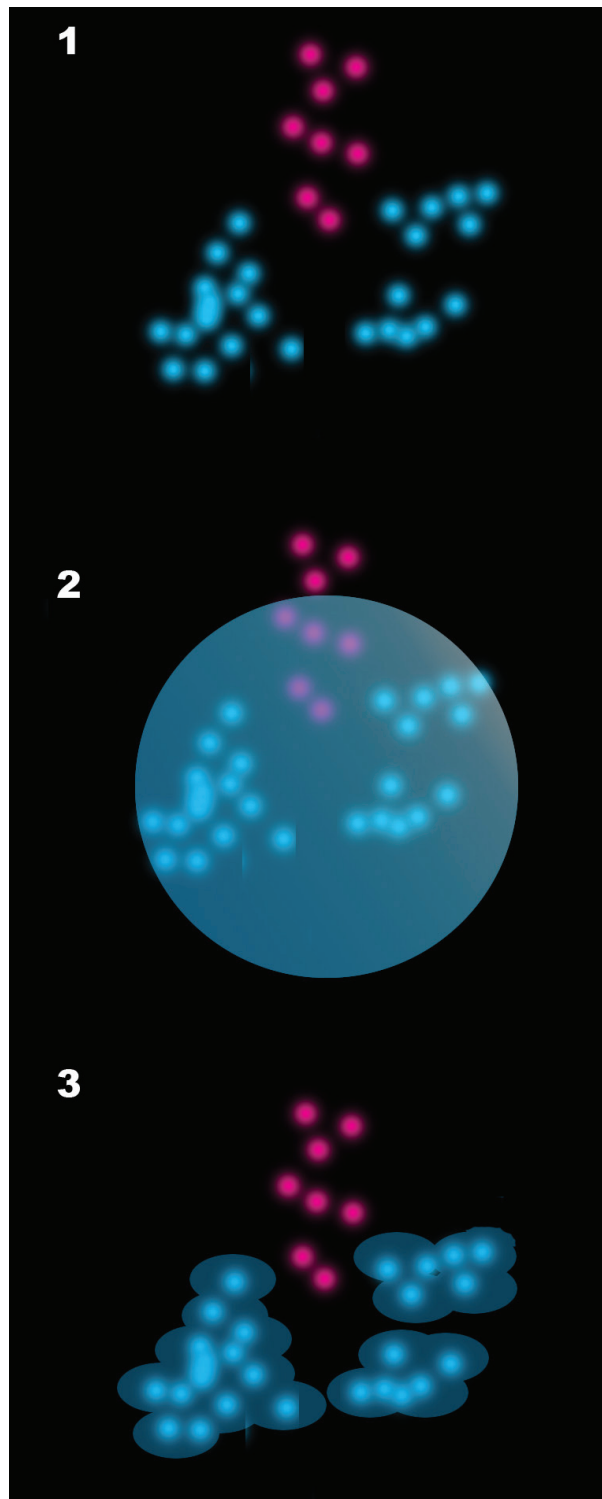


Figure 3 - This figure demonstrates how ECM can be applied to classify data that are not linearly separable (1). A single sphere decision boundary that encompasses all the control subjects (blue dots) did not fully separate the control subjects from the patients. However, ellipsoid boundaries grown from the data points originated from each control subject can better mold to the dataset (3).

agreement is that there is no single universally optimal feature selection technique (YANG et al., 2005); additionally, the existence of more than one subset of features that discriminates the data equally well (YEUNG et al., 2005) should be considered. We believe that each feature selection strategy has its own niche so it is important to know its idiosyncrasies, when to effectively apply it, and also to be aware of its limitations. For example, the output provided by univariate feature rankings can be more intuitive to understand because it analyzes each feature independently. On the other hand, protein subgroups that could possibly interact can only be detected through multivariate techniques, but requiring far more computation time.

### ECM and disease diagnosis

Using clustered ellipsoids to define a decision boundary can be a conservative approach, but it is still able to classify data that are not linearly separable in the feature space (Figure 3). The ellipsoid method is able to encapsulate and efficiently mold to such data and still generalize as seen in Figure 3. The presented method was optimized right from its beginning to deal with the life science multi-class problems showing a new rationale to the traditional classification approaches. Certainly, to further test and validate the method, samples from various types of cancer should be tested and a very large and diverse control subject group should be mapped.

By mapping what is supposed to be normal, we imitate an artificial immune system which is also trained to be at ease with what is normal, and react to what is “not normal”. The immune system continuously learns, since it would be unlikely to have a prior knowledge of all existing pathologies. The new method points toward a path that combines mass spectrometry with ECM to define a cancer free Euclidian domain based on a hyper ellipsoid cluster boundary. ECM could be interpreted as the geometrical definition of a standard, and the ECM algorithm could possibly be applied to other fields of science to track and map standards in quality control, for example. This could be the basis for a multi-class classifier, offering a fast, initial hypothesis as a first step to narrow the possible solution classes. As a second step, this classifier could be combined with other methods (ex. SVM) to then achieve greater confidence binary classifications.

### Availability and requirements

- Project name: Ellipsoid Clustering Machine
- Project home page: <http://www.dbbm.fiocruz.br/labwim/bioinfoteam/templates/archives/ecm/>
- Operating system(s): Platform independent
- Programming language: Perl 5.8.6
- Other requirements: A viewer such as the Cortona VRML client is necessary to view the interactive 3D models. Cortona can be obtained at <http://www.parallelgraphics.com/products/cortona/>
- License: Creative Commons Attribution-NonCommercial-NoDerivs 2.0 License.

- Mass spectra: the raw dataset, or txt files are also available from authors upon request
- Restrictions for non-academic use apply: licensing is needed.

## Acknowledgements

The authors thank the Fundação Ary Frauzino / Fundação Educacional Charles Darwin, Faperj/Faperj-BBP (Cientista do Nosso Estado), CNPq, PDTIS, the Rio de Janeiro proteomic network, the Fiocruz-Inca cooperative agreement and www.genesisdna.com.br for financial support.

## Bibliographic references

ARDEKANI, A.M.; LIOTTA, L.A.; PETRICOIN, E.F., III. Clinical potential of proteomics in the diagnosis of ovarian cancer. **Expert Review of Molecular Diagnostics**, v.2, p.312-320, 2002.

BJELLQVIST, B. et al. Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J Biochem. Biophys Methods*, v.6, p.317-339, 1982.

CARVALHO, P.C. et al. Detection of potential serum molecular markers for Hodgkin's disease. **Jornal Brasileiro de Patologia e Medicina Laboratorial**, v.41, p.99-103, 2005.

CARVALHO, P.C. et al. Differential protein expression patterns obtained by mass spectrometry can aid in the diagnosis of Hodgkin's disease. **Journal of Experimental Therapeutics & Oncology**, v.6, p.137-145, 2007.

CONRADS, T.P. et al. Proteomic patterns as a diagnostic tool for early-stage cancer: a review of its progress to a clinically relevant tool. **Mol. Diagn.**, v.8, p. 77-85, 2004.

GORG, A. et al. Two-dimensional electrophoresis of proteins in an immobilized pH 4-12 gradient. **Electrophoresis**, v.19, p.1516-1519, 1998

HENZEL, W.J. et al. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. **Proceedings of National Academy of Science**, v.90, p.5011-5015, 1993.

LI, J., et al. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. **Clinical Chemistry**, v.48, p.1296-1304, 2002.

LIU, H.; SADYGOV, R.G.; YATES, J.R., III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. **Analytical Chemistry**, v.76, p.4193-4201, 2004.

MAO, Y. et al. Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. **Journal of Biomedicine and Biotechnology**, p.160-171, 2005.

MEHTA, A.I. et al. Biomarker amplification by serum carrier protein binding. **Disease Markers**, v.19, 1-10, 2003.

NIIJIMA, S.; KUHARA, S. Multiclass molecular cancer classification by kernel subspace methods with effective kernel parameter selection. **Journal Bioinformatics and Computational Biology**, v.3, p.1071-1088, 2005.

QU, Y., et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. **Clinical Chemistry**, v.48, p.1835-1843, 2002.

TROYER, D.A. et al. Promise and challenge: Markers of prostate cancer detection, diagnosis and prognosis. **Disease Markers**, v.20, p.117-128, 2004.

UNLU, M.; MORGAN, M.E.; MINDEN, J.S. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. **Electrophoresis**, v.18, p.2071-2077, 1997.

VAPNIK, V.N. **The nature of statistical learning theory**. New York: Springer-Verlag, 1995.


VON, E.F. et al. Fluorescent dual colour 2D-protein gel electrophoresis for rapid detection of differences in protein pattern with standard image analysis software. **Int J Mol. Med** v.8, p.373-377, 2001.

WEINGARTEN, P. et al. Application of proteomics and protein analysis for biomarker and target finding for immunotherapy. **Methods in Molecular Medicine** v.109, p.155-174, 2005.

WESTERMEIER, R. et al. High-resolution two-dimensional electrophoresis with isoelectric focusing in immobilized pH gradients. **J Biochem. Biophys Methods** v.8, p.321-330, 1983.

XU, R.; WUNSCH, D.C. Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data. **IEEE/ACM. Transactions on Computational Biology and Bioinformatics**, v.4, p.65-77, 2007.

YANG, Y.H.; XIAO, Y.; SEGAL, M.R. Identifying differentially expressed genes from microarray experiments via statistic synthesis. **Bioinformatics**. v.21, p.1084-1093, 2005.

YEUNG, K.Y.; BUMGARNER, R.E.; RAFTERY, A.E. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. **Bioinformatics**. v.21, p.2394-2402, 2005. 

## About the authors

### *Paulo Costa Carvalho*

He holds a degree in engineering from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio) and a Masters in cellular and molecular biology (major in bioinformatics) from the Oswaldo Cruz Institute. During his masters, he was advised by Wim Degraeve and Gilberto Barbosa Domont and applied support vector machines to study mass spectrometry based proteomic profiles of control subjects and Hodgkin's disease patients. His main interests include functional genomics, computational and mass spectrometry based proteomics, artificial intelligence, pattern recognition and grid computing. Currently, he is a PhD candidate advised by Prof. Valmir Carneiro Barbosa at the Systems Engineering and Computer Science Program, COPPE, at the Federal University of Rio de Janeiro. His research is focused on the development of methods to extract knowledge (protein interaction networks and differential proteomic) from Multi dimensional Protein Identification Technology (MudPIT) data, when comparing biological systems in different states.

### *Juliana de Saldanha da Gama Fischer Carvalho*

She holds a degree in chemical engineering from the Pontifical Catholic University (PUC-Rio) and a Master's degree from the Faculty of Medicine from Federal University of Rio de Janeiro (UFRJ) where she was advised by Prof. Maria da Gloria da Costa Carvalho and Dr. Eduardo Marcos Paschoal. During her master, she studied the effects of perillyl alcohol and heat shock (HS) treatment in gene expression of human lung adenocarcinoma cell line (A549). Her results showed that HS modifies the chemotherapeutic effects of perillyl alcohol in the extracellular regulated kinase (ERK) activation pathway of A549 cell line by altering the phosphorylation status of p44/42 and that cellular viability decreases with POH in a dose dependant manner. Currently, she is a PhD candidate at the Chemistry Institute of UFRJ and is advised by Prof. Gilberto Barbosa Domont and Prof. Maria da Gloria da Costa Carvalho. Her interest includes mass spectrometry based proteomics, differential gel electrophoresis and Multi dimensional Protein Identification Technology (MudPIT). Her PhD research aims to study the effects perillyl alcohol over glioblastoma multiform, the most aggressive kind of astrocytoma tumor. Her experimental work comprises research on both cell lines, and serum from patients with astrocytomas that are under clinical trial under treatment with perillyl alcohol by intranasal delivery.