COPPE
UFRJ
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia

COMPUTATIONAL ENVIRONMENT FOR DISCOVERY OF NEW MOLECULES WITH DIAGNOSTIC POTENTIAL

André Ramos Fernandes da Silva

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientadores: Paulo Costa Carvalho
Valmir Carneiro Barbosa

Rio de Janeiro
Dezembro de 2016

COMPUTATIONAL ENVIRONMENT FOR DISCOVERY OF NEW MOLECULES WITH DIAGNOSTIC POTENTIAL

André Ramos Fernandes da Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Valmir Carneiro Barbosa, Ph.D.

Prof. Paulo Costa Carvalho, D.Sc.

Prof. Felipe Maia Galvão França, Ph.D.

Dr. Carlos Medicis Morel, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
DEZEMBRO DE 2016

Silva, André Ramos Fernandes da

Computational environment for discovery of new molecules with diagnostic potential / André Ramos Fernandes da Silva. – Rio de Janeiro: UFRJ/COPPE, 2016.

XIII, 34 p.: il.; 29,7 cm.

Orientadores: Paulo Costa Carvalho
Valmir Carneiro Barbosa

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2016.

Referências Bibliográficas: p. 31-32.

1. Proteômica Computacional. 2. Reconhecimento de Padrões. 3. Espectrometria de Massas. I. Carvalho, Paulo Costa *et al*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Dedico esta dissertação aos meus pais,
Lucas Elias e Vanda.

“Wonder is a kind of desire for knowledge. It is a cause of delight because it carries with it the hope of discovery.”

— Saint Thomas Aquinas, *Summa Theologiae*, I-II. Q32. A8.

Agradecimentos

A Deus, pelo universo inteligível que proclama, sem palavras, a vossa inteligência.

Aos meus orientadores, Dr. Paulo Costa Carvalho e Dr. Valmir Carneiro Barbosa, pela direção, atenção e presença constante ao longo de todo o desenvolvimento desse trabalho.

Aos nossos colaboradores do Institut Pasteur de Montevidéu, Dr. Carlos Batthyany, Dra. Rosario Duran e Alejandro Peña, pela participação ativa no desenvolvimento da metodologia e do DiagnoProt, pelo acesso aos dados de doenças renais, e pelo depósito da patente.

Aos nossos colaboradores do Institut Pasteur de Paris, Dr. Diogo Borges Lima e Dra. Julia Chamot-Rooke, pelo feedback valioso e pela colaboração com o desenvolvimento do DiagnoProt.

Ao Marlon D. M. Santos, pela participação no desenvolvimento do módulo de "Find Gold" do DiagnoProt e pela ajuda nas buscas com o Comet.

À Dra. Priscila F. Aquino pelo acesso aos dados de fungos.

Aos nossos demais colaboradores que também tornaram esse trabalho possível: Dr. Gilberto B. Domont, Dr. Jimmy E. Rodriguez e Juliana C. Leal.

Ao CNPq pela bolsa de mestrado.

Resumo da Dissertação apresentada à COPPE/UFRJ, como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

AMBIENTE COMPUTACIONAL PARA DESCOBERTA DE NOVAS MOLÉCULAS COM POTENCIAL PARA DIAGNÓSTICO

André Ramos Fernandes da Silva

Dezembro/2016

Orientadores: Paulo Costa Carvalho
Valmir Carneiro Barbosa

Programa: Engenharia de Sistemas e Computação

A proteômica é uma ciência multidisciplinar que realiza o estudo em larga escala de proteínas. Algumas de suas aplicações são identificação e sequenciamento de proteínas, quantificação, e identificação de interações proteína-proteína. Para realizar essas tarefas, a proteômica faz extenso uso de espectrometria de massas e de inteligência artificial.

Todavia, em média 75% dos espectros de massas analisados não são identificados pelas estratégias de busca consideradas estado-da-arte. Esse trabalho apresenta metodologia inovadora, capaz de listar espectros de alta qualidade que são discriminativos entre diferentes condições biológicas independentemente de serem identificados pelas ferramentas de busca existentes. Exemplificamos uma aplicação desta metodologia ao listar espectros discriminativos entre três espécies de fungos do gênero *Aspergillus*. Adicionalmente, nossa metodologia mostrou-se capaz de discriminar entre amostras de pacientes diagnosticados com as doenças renais nefropatia lúpica e nefropatia por IgA.

A metodologia foi implementada no software DiagnoProt, o qual está disponível para download no site http://www.patternlabforproteomics.org/diagnoprot/.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

COMPUTATIONAL ENVIRONMENT FOR DISCOVERY OF NEW MOLECULES WITH DIAGNOSTIC POTENTIAL

André Ramos Fernandes da Silva

December/2016

Advisors: Paulo Costa Carvalho
Valmir Carneiro Barbosa

Department: Systems Engineering and Computer Science

Proteomics is a multidisciplinary science that performs large-scale study of proteins. Some of its applications are identification and sequencing of proteins, quantification, and identification of protein-protein interactions. To accomplish these goals, proteomics relies on intense use of mass spectrometry and artificial intelligence.

On average, 75% of all mass spectra are not identified by state-of-the-art search strategies. In this work, we introduce a methodology for listing high-quality spectra that are discriminative among different biological conditions, independently of being identified by existing search tools. We exemplify the usefulness of our methodology by shortlisting discriminative spectra among three types of fungi from the *Aspergillus* genus. Moreover, our methodology could discriminate between samples of patients with the kidney diseases lupus nephropathy versus IgA nephropathy.

The methodology was implemented in a software termed DiagnoProt that is available at http://www.patternlabforproteomics.org/diagnoprot/.

Summary

List of Figures

List of Tables

List of abbreviations, symbols and unities

ACN	Acetonitrile
CID	Collision-induced dissociation
CNI	Cumulative normalized intensity
DDA	Data-dependent acquisition
IgAN	Immunoglobulin-A-Nephropathy
KB	Knowledge base
LC	Liquid chromatography
LC-MS/MS	Liquid chromatography online with tandem mass spectrometry
LMW	Low molecular weight
LN	Lupus Nephritis
LOOCV	Leave one out cross-validation
LTQ	Linear Trap Quadrupole
MS	Mass spectrum; mass spectrometry
MS1	Survey scan (usually refers to a peptide m/z profile)
MS2 or MS/MS	Mass spectrum of dissociated ions
m/z	Mass to charge ratio
TIC	Total ion current
PCA	Principal Component Analysis
PKB	Knowledge base file extension
PTM	Post-Translational modification

1 Introduction

1.1 Proteomics

Proteomics is a multidisciplinary science; its goal is to perform large-scale analysis of proteins in complex biological samples. Some of its applications are identification and sequencing of proteins, quantification, and identification of protein-protein interactions. It revolutionized research in the fields of biology, biotechnology, and medicine. Today it is practically inconceivable to characterize organisms or study pathologies without considering proteomics. To achieve its goals, proteomics makes intense use of mass spectrometers and artificial intelligence (CARVALHO; BARBOSA, 2010).

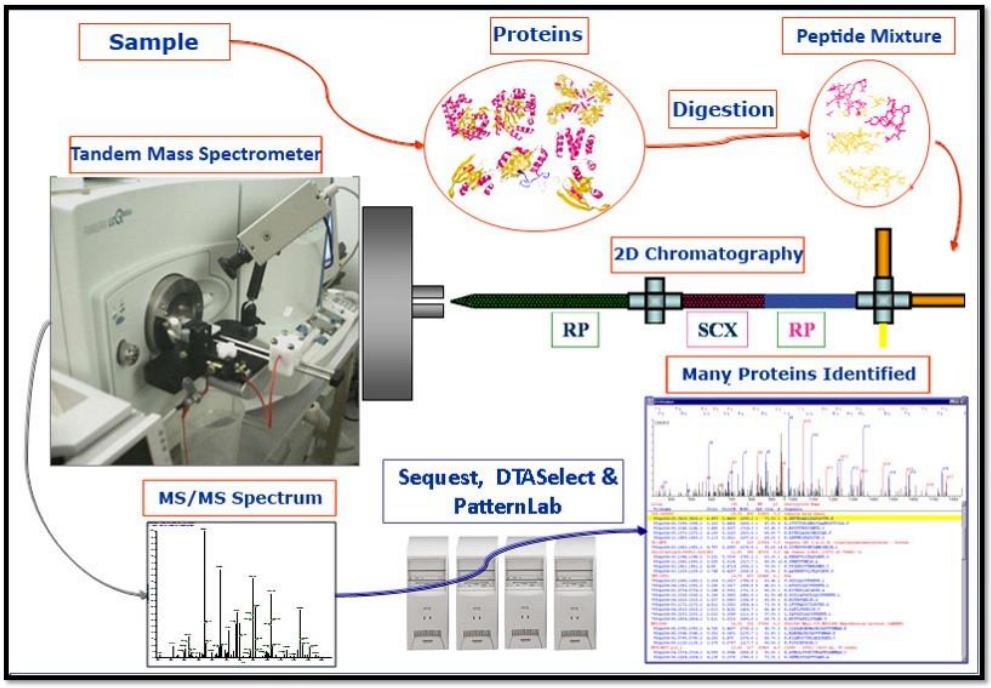


Figure 1 *Example of a proteomics workflow.*[1]

Figure 1 shows an example of a proteomic workflow. A biological sample is prepared in order to extract the proteins. Subsequently, proteins are digested into peptides using a protease (e.g. Trypsin) and further separated using liquid chromatography (LC) online with the tandem mass spectrometry. The peptides are ionized in a process called electrospray (FENN et al., 1989) as they are "injected" into the spectrometer and ultimately analyzed to produce mass spectra. A mass spectrum is usually represented as a 2-dimentional graph; the x-axis shows the mass to charge ratios (m/z) of the ions, and

[1] Figure from (CARVALHO; BARBOSA, 2010)

the y-axis shows their respective intensities. A mass spectrum acquired from the molecules eluting from the column is usually referred to as an MS1 spectra. The mass spectrometer can isolate and dissociate molecules within a very narrow m/z range (e.g., +-1 m/z); hence, generating an MS2 spectrum of such precursor ions. Thus, a MS2 spectrum is a spectrum originating from the analysis of fragment ions. A typical proteomic experiment generates hundreds of thousands of spectra. Therefore, it is unfeasible to attempt to analyze proteomic data without specialized software.

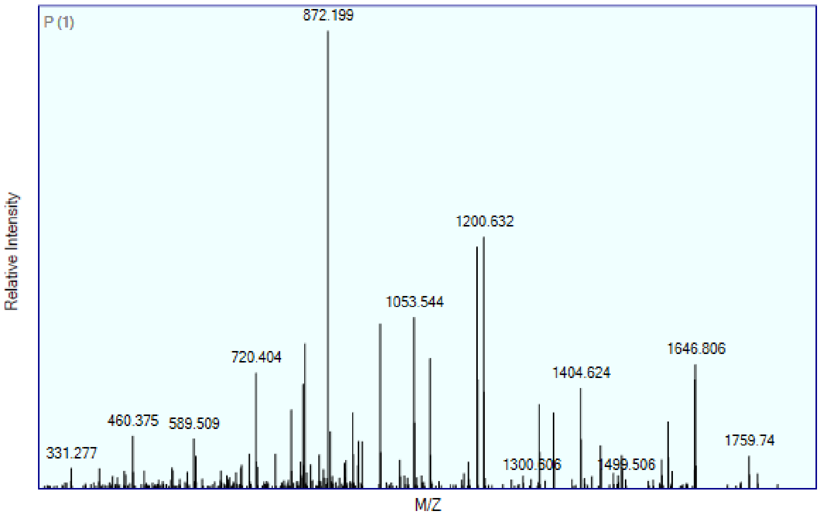


Figure 2 *Example of a MS2 mass spectrum.*

1.2 Peptide Spectrum Matching

Peptide Spectrum Matching (PSM) is a strategy to identify peptides and proteins analyzed with a mass spectrometer. A general PSM workflow is illustrated in Figure 3. Experimental tandem mass spectra (MS2) (i.e., spectra of dissociated peptides) are generated by the mass spectrometry and stored in a computer file (e.g. RAW). The software scans a protein sequence database and may perform an *in silico* digestion to compute a list of putative peptides. For each experimental spectrum, the PSM approach shortlists peptide sequences, within a mass tolerance compatible to the equipment at hand, and then computes scores, for each peptide assignment, that compare the similarity between the theoretical versus the experimental mass spectrum. A widely adopted score is the cross-correlation (Xcorr) (ENG et al., 1994). The top highest ranking peptides are reported for each experimental mass spectrum and saved in a result file (e.g. SQT file). The result file must be later analyzed by a PSM post-processing software (e.g. SEPro) that will single out unreliable assignments according to statistical tests (CARVALHO et al., 2012).

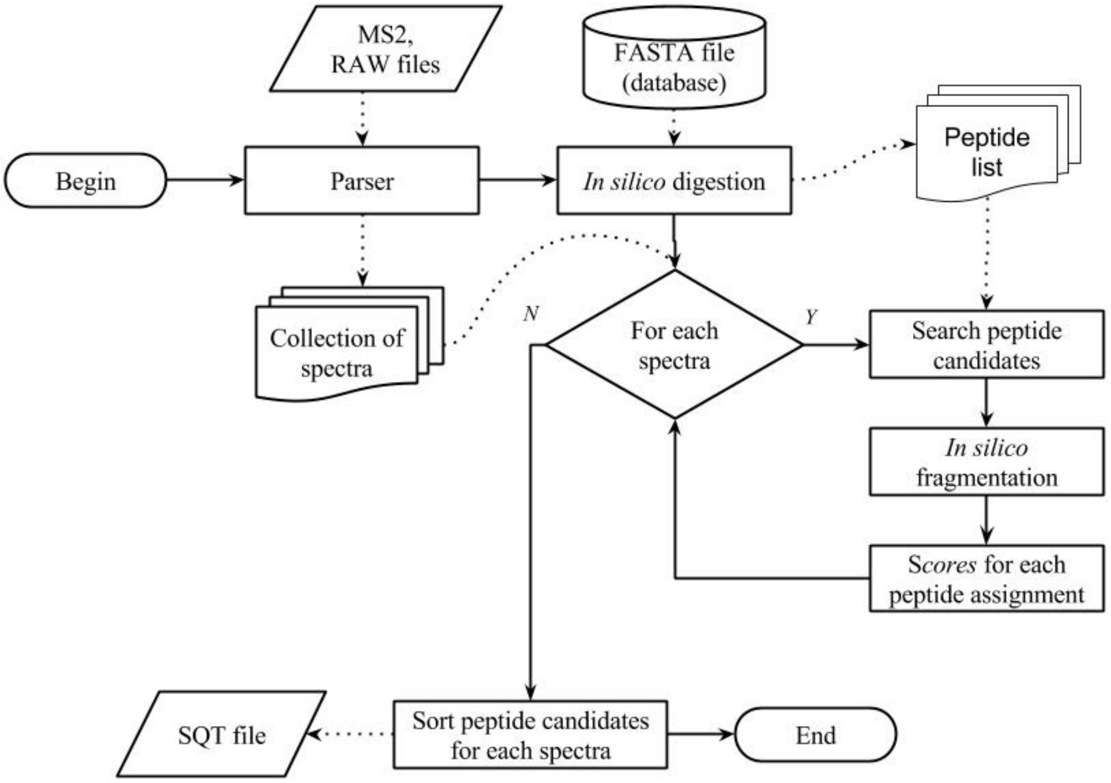


Figure 3 *PSM workflow.*

PSM is considered the gold standard of proteomic search because it is more accurate and sensitive than methods that assume no knowledge of sequenced proteins (LIMA, DIOGO BORGES et al., 2013). However, it has some limitations. It considers many putative candidates that probably will never be seen in actual data. An aggravating factor of a big peptide search space is the loss of sensitivity (BORGES et al., 2013). In general, post-translational modifications (PTMs) must be specified a priory. It only identifies proteins in the database or, in some cases, with one amino acid difference in some error tolerant searches implementations.

1.3 Identifying microorganisms with commercial solutions

A trend that recently emerged is to diagnose microorganisms with mass spectrometry. This approach relies on a completely different paradigm as compared to PSM. In this approach, a single mass spectrum, referred to as a protein fingerprint, is compared to those previously obtained from known organisms and stored in a spectrum database. These approaches require growing the bacteria to be identified in a culture on a petri dish, enriching the sample for proteins (e.g., metal binding proteins), and obtaining a mass spectrum of the protein profile of this sample. A commercial example of this application is the MALDI Biotyper, from Bruker (https://www.bruker.com/products/mass-spectrometry-and-separations/maldi-biotyper/overview.html). A typical workflow is illustrated in Figure 4.

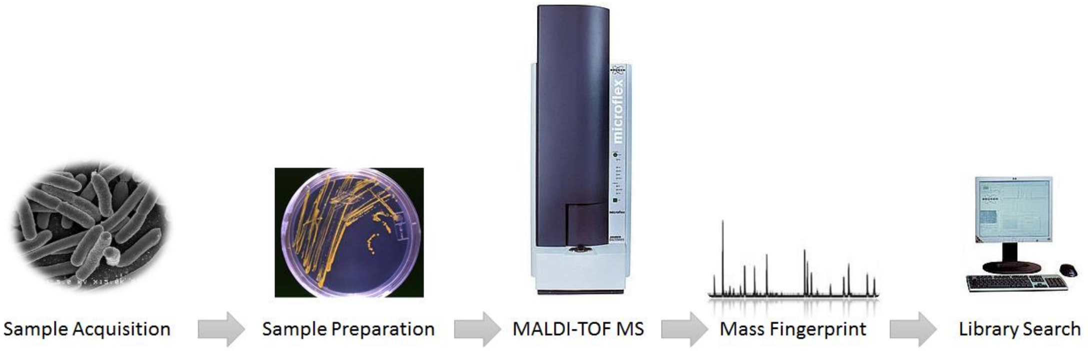


Figure 4 *Identification of microorganisms with MALDI-TOF.*

Although this approach has proven effective in the task of identification of many microorganisms, there are several limitations (WELKER, 2011). These limitation stem, in part, when trying to classify organisms that share vary similar proteomes. As such, these organisms cannot be separated relying only within the information provided within a single mass spectrum. Some examples of when this strategy fails are seen when classifying between *E. coli* vs *Shigella* or bacteria that are resistant or not to an antibiotic. Moreover, it is incapable of classifying pathological states when taking, for example biological fluid samples (e.g. urine). Another aggravating limitation of this approach is that it cannot identify the proteins with the single spectrum mass fingerprint. Therefore, the diagnosis relies only on mass spectral peaks; no biological information, that can help disclose mechanisms of the pathology at hand, at a molecular level, are provided by the method.

2 Motivation

A recent study showed that, on average, 75% of mass spectra remain unidentified by state-of-the-art proteomic search engines (GRISS et al., 2016). Search strategies that rely on protein sequence databases are unable to identify proteins that have not been characterized before (i.e., included in the sequence database). In addition, they are limited in only identifying mutations and post-translational modifications (PTMs) that are known or have been specified a priori. On the other hand, *de novo* sequencing strategies, that do not rely on sequence databases can, in some cases, overcome some of these shortcomings; yet, their sensitivity for this task is significantly lower than PSM as there are no bearings provided by previous knowledge, the increased set of possibilities results in a loss of sensitivity, finally, PTMs must be specified *a priori* (BORGES et al., 2013).

Ever since the "post-genomic age", computing power are increasingly necessary for research in biological sciences (JAMES, 1997). To further develop a proteomic application that can discriminate between biological conditions and even aid diagnosis of pathologies, a computational methodology must be able to efficiently handle these spectra that remained unidentified because these mutations and PTMs could be discriminative evidence between different conditions. Our approach introduces a new paradigm on how microorganisms are classified or pathologies are diagnosed by considering a collection of tandem mass spectra. This spectral library contains sequence information and the depth necessary to discriminate between samples that are proteomically alike; yet, lifts the shortcomings of having to identify these molecules as a first step.

3 Objectives

The development of a computational environment that can organize and explore mass spectra of different biological conditions. The methodology is to be implemented in a user-friendly software that can:

- Efficiently handle millions of spectra and save them into an easily sharable data structure called the knowledge base;
- Communicate with existing proteomic search engines (e.g. SEQUEST, Comet, ProLuCID, etc.);
- Shortlist spectra that can discriminate among different biological conditions and pinpoint high-quality unidentified spectra among these;
- Graphically represent distances among existing biological conditions and biological samples of the knowledge base;
- Classify unknown spectral profiles based of pair-wise spectral comparison between the conditions in the knowledge base.

4 Methodology

4.1 Overview of the shortlisting discriminative molecules procedure

Figure 5 shows the general overview of the shortlisting method. A collection of MS2 spectra originating from a biological condition is filtered by the quality control (QC), which is an artificial intelligence filter that removes noisy spectra and spectra that are not very likely to be representative of meaningful molecules (e.g. contaminants and spectra that did not fragment well). The selected spectra are clustered by a spectral similarity function in order to eliminate redundancy, and then are inserted into a data structure called knowledge base (KB). Every biological condition will have its own collection of spectral clusters in the KB. The discriminating analysis finds clusters that only occur in one single condition. A PSM search is performed in the discriminative clusters, and those who rank very poorly are shortlisted as the high-quality discriminating spectra that could not be identified. Since they passed QC and are discriminative, but are not identified, they are likely to be spectra related to discriminating PTMs and/or mutations indicative of that biological condition. In the following sections, we provide detailed explanation of how these modules work.

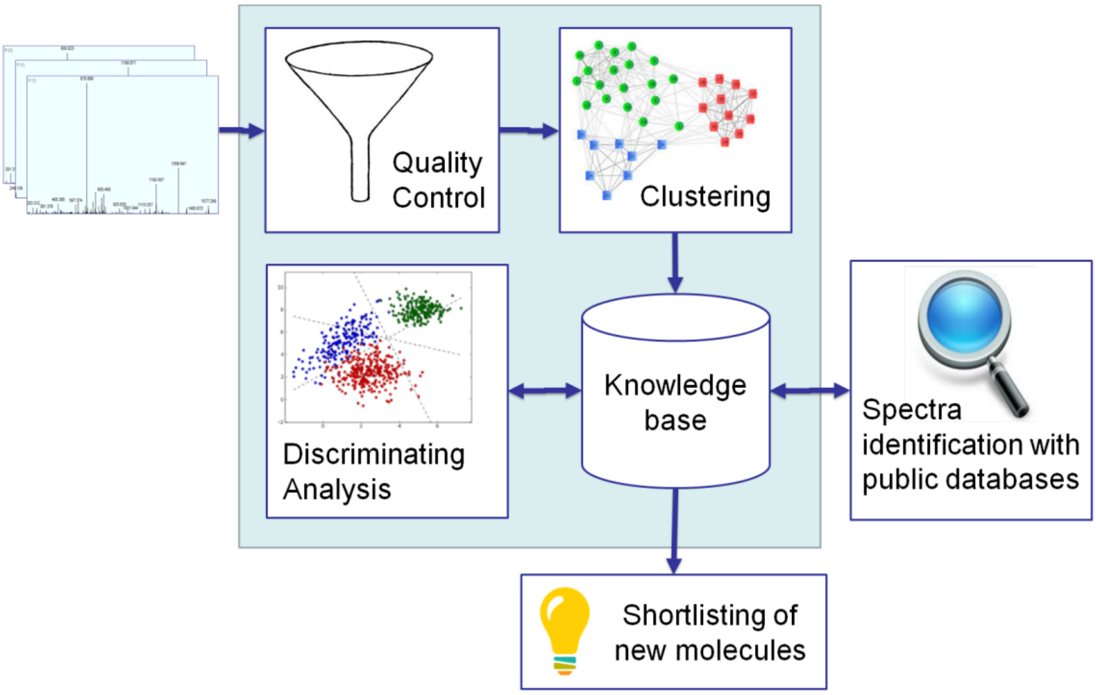

Quality Control
Clustering
Knowledge base
Discriminating Analysis
Spectra identification with public databases
Shortlisting of new molecules

Figure 5 *Shortlisting high-quality unidentified molecules.*

4.2 Spectral Quality Control

The QC is implemented as seven quality filters: precursor charge state filter, minimum number of peaks (Min. No. Peaks), minimum retention time (Min. Ret. Time), minimum relativize intensity (Min. Rel. Intensity), minimum Xrea (Min. Xrea), maximum Balance (Max. Balance) and minimum spectral count (Min. Spec. Count).

4.2.1 Precursor charge filter

Most of spectra with precursor charge state 1+ are contaminants that are not peptides and therefore are disregarded.

4.2.2 Minimum number of peaks

A low number of peaks might be an indicator that the precursor ion did not fragment well, and in this case the spectrum would not contribute to discriminative information and should be disregarded. If the number of peaks of the spectrum is less than Min. No. Peaks, then it is kept out of the KB.

4.2.3 Minimum retention time

A minimum retention time may also be used to filter data. Spectra generated before the value set in this parameter are disregarded. The appropriate setting for this parameter, depends on the equipment, laboratory conditions and samples being processed. Similarly, the acetonitrile-to-water gradient used in the mass spectrometer may also affect this parameter. For example, a setting of 10.00 minutes may be appropriate for a process going from 5% to 40% acetonitrile (ACN) over two hours. If the process applies a gradient of 5% to 40% ACN over one hour, a minimum retention time of 5.00 minutes may be more appropriate. The motivation is that peptides will, in most cases, only elute from the chromatographic column after a certain percentage of ACN is achieved.

4.2.4 Minimum relative intensity

The minimum relative intensity removes noisy low intensity peaks from a spectrum. Let the relative intensity ρ, in relation to the most intense peak of a spectrum, be defined as follows.

$$\rho = \frac{\text{peak intensity}}{\text{highest intensity}}$$

All peaks that have $\rho < \text{Min. Rel. Intensity}$ are removed from the spectrum.

4.2.5 Minimum Xrea

Xrea is a signal-to-noise score proposed by (NA; PAEK, 2006). To calculate the Xrea of a spectrum, it is necessary sort all the peaks of the spectra in ascending order. This is called relative intensities curve. Let n be the number of peaks in a spectrum. Let

$rank(i)$ be the rank of peak i in the relative intensities curve, the rank of the least intense peak be 1, and the rank of the most intense peak be n. Let $I_{raw}(i)$ be the measured raw intensity of peak i. Let TIC be the total ion current of the spectrum, that is, the sum of all raw intensities of its peaks.

$$TIC = \sum_{\forall i} I_{raw}(i)$$

From the curve of relative intensities, we compute a curve of cumulative normalized intensities. Let $CNI(i)$ be the cumulative normalized intensity of peak i.

$$CNI(i) = \sum_{\forall x \mid rank(x) \leq rank(i)} \frac{I_{raw}(x)}{TIC}$$

That is, $CNI(i)$ is the sum of all intensities of peaks whose intensities are less than or equal the intensity of peak i. Figure 6 shows examples of CNI curves for four spectra of different qualities. Xrea is defined as follows.

$$\text{Xrea} = \frac{\text{area XX}}{\text{triangle area} + \alpha}$$

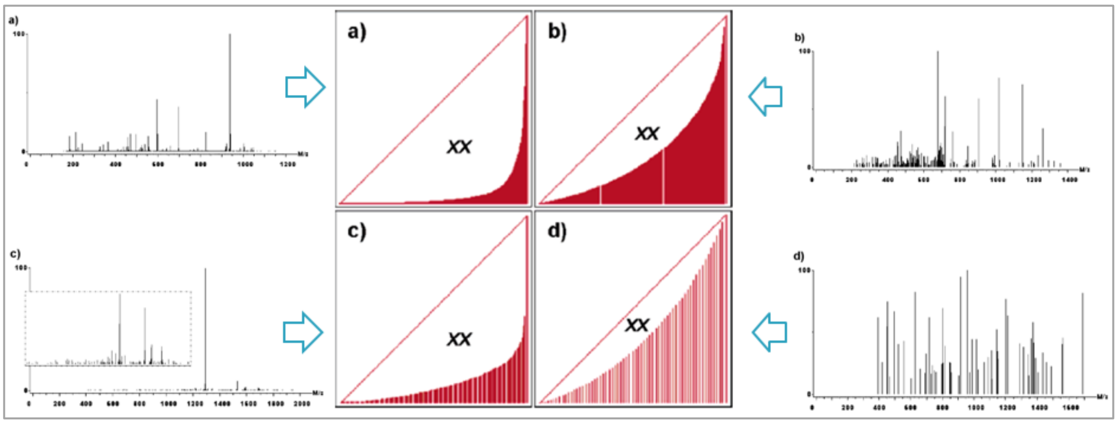


Figure 6 *CNI curves for spectra of different qualities.*[2]

The area XX is the area between the CNI curve and the diagonal of the triangle. The spectra considered to be good for identification are those with few intense peaks distributed along the spectrum (see spectrum **a** in Figure 6). On the other hand, low quality spectra lots of peaks with close intensities, therefore making it very difficult to discriminate true fragment ion peaks than noise peaks (see spectrum **d** in Figure 6). By this definition, spectrum a will have an excellent Xrea, while d will have a very poor score, and spectrum d will be something in between. However, there are some spectra (possibly those that do not fragment well) that have one very intense peak much greater

[2] Figure adapted from (NA; PAEK, 2006)

than any other peak (see spectrum **c** in Figure 6). These spectra will have a high area XX, but are bad for classification nonetheless. The penalty factor α in the equation prevents such spectra to have a high Xrea. The term α is defined as the $I_{raw}(h)/TIC$, where h is the most intense peak of the spectrum.

A minimum Xrea threshold can be set. All spectra with Xrea below the threshold will be disregarded during search. The authors have determined empirically that a useful threshold could be $0.3 \leq \text{Min. Xrea} \leq 0.4$. DiagnoProt's default value is $\text{Min. Xrea} = 0.4$, but the user can fine-tune it to their needs.

4.2.6 Maximum Balance

The Balance is quality score that is complementary to Xrea. It is important to have a complementary score, because Xrea alone does not guarantee a favorable distribution of peak intensities along the spectrum. While Xrea is a signal-to-noise score, the Balance measures how the peak intensity distribution of a spectrum diverges from a model distribution, and thus favors an overall intensity distribution along the spectrum.

Here we describe how a model distribution is estimated from a set of high-quality spectra. First, we bin all the spectra in 13 bins using 100 Da bin size, 200 Da minimum bin m/z, 1500 Da maximum bin m/z and 0 bin offset.[3] Let $\vec{x} \in \mathbb{R}^{13}$ be a vector of intensities representing a binned spectrum with the given parameters, $\vec{x} = (x_1, \cdots, x_{13})$. Let $TIC(\vec{x})$ be the sum of all binned intensities of $\vec{x}$.

$$TIC(\vec{x}) = \sum_{i=1}^{13} x_i$$

Thus, we estimate a model intensity distribution $\hat{G}$ from a dataset of high quality spectra as follows.

$$\hat{G} = (g_1, \cdots, g_{13}) = \frac{\sum_{\forall \vec{x}} \vec{x}}{\sum_{\forall \vec{x}} TIC(\vec{x})}$$

It is easy to show that $\hat{G}$ is a vector that $\sum_{i=1}^{13} g_i = 1$.

Once the model distribution is chosen, the Balance score is calculated as the Kullback-Leibler divergence (KL-divergence) between the intensities of the spectrum and the model distribution. Let m be a mass spectrum and $B(m) = (b_1, \cdots, b_{13}) = Bin(m)/TIC\big(Bin(m)\big)$ be the intensities vector of binned and normalized spectrum. It

[3] Please, see Binning in section 4.3 for more details.

is also easy to show that $\sum_{i=1}^{13} b_i = 1$. Let z be the charge state of m's precursor ion. Then, $\mathrm{Balance}(m)$, the Balance score of $m,$ is defined as follows.

$$\mathrm{Balance}(m) = D_{KL}\big(B(m)||\hat{G}\big)$$

$D_{KL}(P||Q)$ is the KL-divergence between distributions P and Q.

$$D_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$$

The higher the KL-divergence, the more the two distributions diverge from each other. Thus, Balance is a measure of how a spectrum diverges from a model distribution of high-quality spectra. The default Max. Balance is set to 1, but users can set a threshold more appropriate to their needs.

4.2.7 Minimum Spectral Count

A spectral cluster, as defined before, is a spectrum that is elected to be representative of a set of similar spectra. If this set has less spectra than Min. Spec. Count, then it is disregarded during the search. The intuition behind this is that spectra that have rare occurrence have dubious discriminative power, and therefore should be kept out of the analysis. For example, suppose a condition has a spectral cluster with only one spectra, a singleton set, and this spectrum has only been observed in that condition. It is obviously uncertain whether this spectrum could be regarded as discriminative or is just there by chance, because it was not consistently observed in other replicates.

4.3 Binning

Binning is a spectral transformation that is performed after QC and before clustering. Four parameters control the transformation: Min. Bin m/z, Max. Bin m/z, Bin Size and Bin Offset.

Peak intensities measured at an m/z that is less than the Min. Bin m/z, or greater than the Max. Bin m/z setting, may be disregarded or set to 0. Under standard lab settings, using ordinary lab equipment, the Min Bin m/z setting may be set to 200.00, and the Max Bin m/z setting may be set to 1700.00. These data filter values are chosen because 95 % of intensity measurements fall within these m/z ranges. However, practitioners using different equipment or lab settings, or analyzing samples or using enzymes that are expected to generate intensities at greater than 1700.00 m/z or less than 200.00 m/z may change Min & Max Bin m/z parameters to values that better conform to the data they are analyzing.

A bin is defined as a m/z interval $[l, u[$, where l is the m/z lower bound and u is the m/z upper bound. Let $mz(i)$ be the m/z of peak i. The bin intensity $I_{bin}(l, u)$ is defined as the sum of all peak intensities whose m/z fall into the bin m/z interval, that is,

$$I_{bin}(l, u) = \sum_{l \leq mz(i) < u} I_{raw}(i)$$

The binning procedure also have Bin Size (s) and Bin Offset (o) parameters. For all u and l, Bin Size is $u - l$. A Bin Size of 1.0005 may be used. This size is selected based on standard lab and equipment settings and depends on the resolution on the mass spectrometer.

Bin Offset is defined as the lower bound of the first bin. Therefore, the first bin will be $[o, o + s[$, the second bin will be $[o + s, o + 2s[$, and the ith bin will be $[o + (i - 1)s, o + is[$. This offset may be applied to help distinguish between amino acid combinations. It is preferably set to 0.40, because no combination of sums of amino acid masses will coincide with the sum of the offset plus a multiple of Bin Size. As discussed above, the ideal data parameters used will vary depending on the equipment used. When using state-of-the-art high resolution equipment, no offset may be necessary.

Thus, a binned spectrum is a transformation performed in a mass spectrum such that, the m/z of the ith peak is $o + (i - 1)s$ and its intensity is $I_{bin}(o + (i - 1)s, o + is)$. All bins with zero intensity can be discarded to same memory. For simplicity of notation, we write the intensity of the ith bin as $I_{bin}(i) = I_{bin}(o + (i - 1)s, o + is)$. Once all bin intensities and m/z are calculated, the bin intensities are divided by $\sqrt{\sum_i I_{bin}(i)^2}$ so that the vector of intensities becomes a normalized vector with norm 1.

4.4 Clustering

A spectral cluster is a spectrum elected to be representative of a set of similar spectra. The objective of this procedure is to reduce redundancy in the KB. For example, Figure 7 shows two very similar spectra that clearly represent the same fragmentation pattern. Three KB creation parameters control the clustering: Similarity Threshold, Precursor Tolerance, and Ret. Time Tolerance.

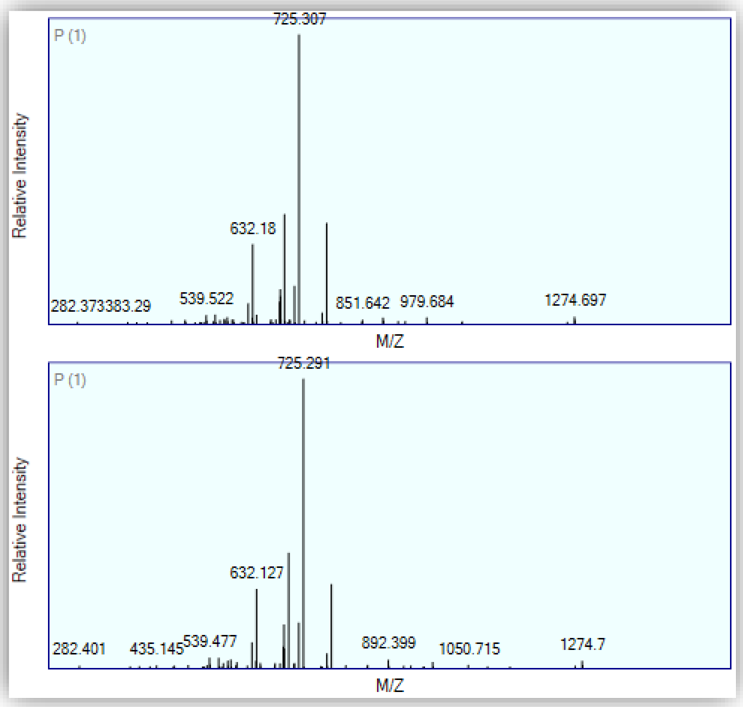


Figure 7 *Example of two very similar spectra.*

The Precursor Tolerance is the maximum allowed absolute difference between the precursors m/z of the spectra. All spectra with higher precursor m/z difference will not be clustered.

The Ret. Time Tolerance is the maximum allowed absolute difference between chromatography retention time of the spectra. All spectra with higher retention time difference will not be clustered. This parameter is optional. If a user does not want to use this parameter, the retention time will not be considered during clustering.

The Similarity Threshold is the minimum similarity score allowed for two spectra to be considered similar. The similarity score used in this methodology is the dot product of normalized binned intensities. Let $\vec{x} = (x_1, \cdots, x_n)$ and $\vec{y} = (y_1, \cdots, y_n)$ be the normalized binned intensities vectors of two spectra, and let $\vec{x}.\vec{y} = \sum_{i=1}^{n} x_i y_i$ be the intensities dot product of intensities. All spectra that have $\vec{x}.\vec{y} <$ Similarity Threshold will not be clustered.

Since $\vec{x}.\vec{y}$ demands many float-point number computations, DiagnoProt performs Ret. Time Tolerance and Precursor Tolerance before computing the dot product. If a pair of spectrum does not pass the first tests, then computation time is saved. Another important test that is performed before $\vec{x}.\vec{y}$ is the binned base peak test. The base peak of a spectrum is its most intense peak. Since the spectra are binned, all the peaks have the same corresponding m/z values. If the m/z of the base peak of the two spectra are different, then the two spectra will not be clustered.

Two spectra are similar if, and only if, they pass all the clustering tests. When two spectra are clustered together, the spectrum with the highest Xrea is chosen to be the representative of the cluster and the other is discarded. Given a collection of binned spectra, the clustering procedure finds all the spectral clusters and stops when there are no more spectra that can be clustered. All cluster collections are saved into their respective biological condition in the KB.

4.5 Knowledge Base

The knowledge base is a data structure that stores information about the experiment, the values of the parameters (e.g., QC parameters, thresholds, etc.), and the biological conditions with their respective spectral clusters.

4.6 Discriminating Analysis

When the KB is created, it is possible to search for spectral clusters that are discriminative among biological condition. There must be with at least two biological conditions in the KB. The discriminating analysis works by comparing each pair of biological conditions, and searching similar spectral clusters that occurs in both conditions, and the ones that are found not to be similar in the other condition.

A cluster m belonging to biological condition A occurs in biological condition B if, and only if, m is similar to a cluster belonging to B. A cluster m is exclusive of a biological condition A if, and only if, it occurs only in A. Thus, the discriminating analysis is a search for the exclusive of each biological condition using the same similarity relation defined for comparing spectra for clustering.

4.7 Shortlisting of new molecules

As show by (GRISS et al., 2016), the majority of spectra in a proteomic experiment remain unidentified by proteomic search engines. All the steps previously described for creating the KB and performing a discriminating analysis do not demand that a spectrum is identified, it only demands that the spectrum pass the QC filters. However, a user my attempt to identify the spectral clusters in the KB, specially the discriminative clusters. The shortlisting of new molecules works by finding those clusters that could not be identified by Comet (ENG et al., 2013, 2015). A very low Xcorr threshold must be given (e.g. < 1.5) so that a peptide (or protein) assignment is regarded as an unreliable match. A Comet search is then performed in these discriminative spectra, and those bellow the Xcorr threshold are shortlisted as new molecules with discriminating power. These high-quality spectra that could not be identified could be related to mutations or PTMs that are representative of their respective biological conditions.

4.8 Spectral Profile Classifier

The spectral profile classifier performs identification of samples of unknown biological conditions by pair-wise comparisons of spectral cluster collections. Figure 8 illustrates the general workflow of the classifier. A set of spectra belonging to an unknown biological condition is filtered using QC and clustered using the same procedure and parameters values set for KB creation. Let U be the collection of spectral clusters from this unknown sample, and C_i the collection of spectral clusters in the ith condition in the KB. Then, a similarity score between U and C_i is computed as the Jaccard index $J(U, C_i)$. The Jaccard index between two collections of spectral clusters A and B is defined as $J(A, B) = |A \cap B|/|A \cup B|$. In the context of this application, $|A \cap B|$ is defined as the number of spectra in A that occurs[4] in B, and $|A \cup B|$ is defined as $|A| + |B| - |A \cap B|$. The classifier assigns to U the condition that maximizes $J(U, C_i)$.

$$\text{Condition}(U) = \underset{i}{\text{argmax}}\, J(U, C_i)$$

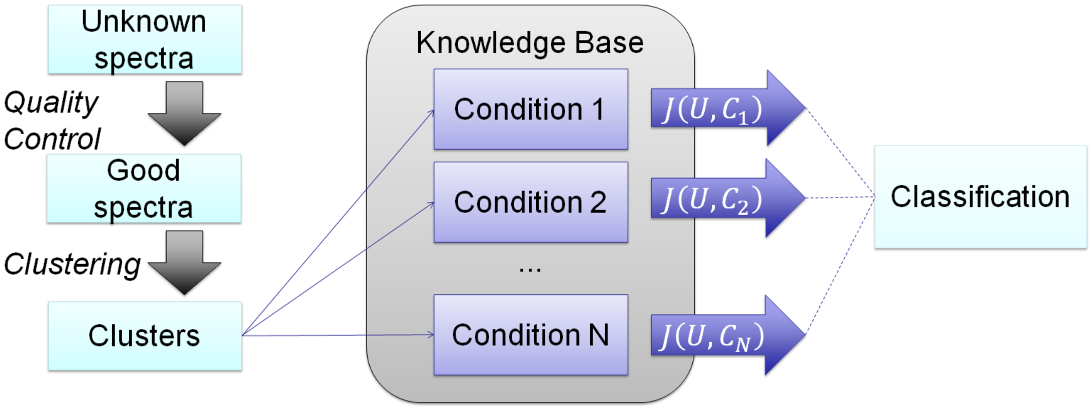


Figure 8 *Spectral profile classifier.*

A k-fold cross-validation can be performed to give an estimate of the out-of-sample error (ABU-MOSTAFA et al., 2012). This method uses a leave one out cross-validation (LOOCV) procedure. A LOOCV estimates classification error by creating multiple

[4] See section 4.6 for the definition of spectral cluster occurrence.

KBs, one for each biological sample in the dataset. A KB is created leaving the validation sample out of the KB. The spectral profile classifier then analyzes this left-out sample. Since we know to which biological condition the sample belongs to, but the classifier does not, we compute if the condition assigned by the classifier corresponds to the true condition of the sample. In this way, the classification error is estimated by computing all LOOCV condition assignments and calculating the overall success rate.

4.9 Kidney diseases dataset

Urine samples of patients diagnosed by renal biopsy with Lupus Nephritis (LN) and Immunoglobulin-A-Nephropathy (IgAN) were collected from spontaneously voided midstream of the second morning urination and stored at 4ºC until preparation. Consent was obtained from all apparently healthy individuals and patients. Low molecular weight (LMW) fractions were digested with trypsin in solution after reduction and alkylation of cysteines. Subsequently, LMW fractions were desalted using OMIX RP micro-columns and loaded onto a nanoLC system (EASY-nLC 1000, Thermo Fischer Scientific). Eluted molecules were detected in an LTQ Velos instrument (Thermo Fischer Scientific) using a data-dependent acquisition (DDA) mode with a dynamic exclusion list. This dataset has 2 biological conditions, 12 biological samples from LN patients and 11 from IgAN patients, resulting in 23 RAW files. The analysis and data generation was performed at the Analytical Biochemistry and Proteomics Unit at Institute Pasteur Montevideo.

4.10 Fungi dataset

A fungi dataset was obtained from three species of fungi of the *Aspergillus* genus: *Aspergillus flavus*, *Aspergillus oryzae* and *Aspergillus parasiticus*. The samples were prepared according to protocols described in (AQUINO et al., 2012). Tryptic peptides were loaded into a nanochromatography system (Easy-nLC II, Proxeon), and eluted directly into an LTQ Orbitrap Velos (Thermo Fischer Scientific). Spectra acquisition was performed using DDA, automatically alternating between full scan MS and MS/MS. The top 10 most intense ions, with charge $\geq 2+$ were isolated and fragmented by collision-induced dissociation (CID) using normalized collision energy 35. This dataset has 4 conditions: one for each fungi species and one control condition. Each condition has 2 biological replicate, and each biological replicate has 3 technical replicates, resulting in 24 RAW files. The analysis was done at the Proteomics Unit and Biochemistry Department at Federal University of Rio de Janeiro.

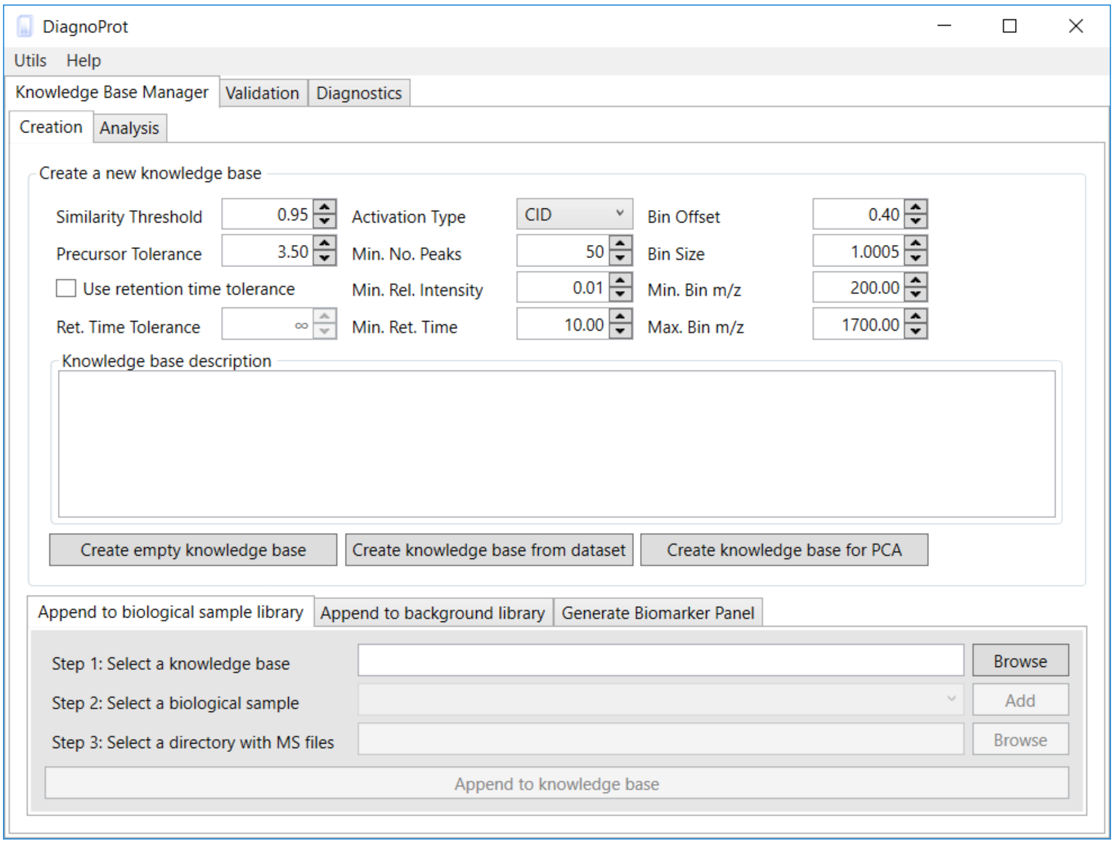
5 Results

5.1 Implementation

All the methodology was implemented in a software, DiagnoProt, that is available at http://patternlabforproteomics.org/diagnoprot/. DiagnoProt was implemented in C#, using .NET framework 4.5. It references PatternLab (CARVALHO et al., 2015) modules for MS files parsers and spectral view controls; Comet for PSM search; Google Protocol Buffers for serialization; DotNetZip for handling ZIP files; Accord.net for PCA; and OxyPlot for charts.

5.2 Knowledge Base Manager

The Knowledge Base Manager makes it easy to configure, create and manage KBs. Figure 9 shows the DiagnoProt's KB Manager interface. During KB creation, it is possible to define parameters that affect QC, binning, and clustering. A user can either create an empty KB or create a KB from a dataset of MS files.


DiagnoProt
Utils
Help
Knowledge Base Manager
Validation
Diagnostics
Creation
Analysis
Create empty knowledge base
Create knowledge base from dataset
Create knowledge base for PCA
Append to biological sample library
Append to background library
Generate Biomarker Panel
Browse
Add
Browse
Append to knowledge base

Create a new knowledge base

Similarity Threshold	0.95	Activation Type	CID	Bin Offset	0.40
Precursor Tolerance	3.50	Min. No. Peaks	50	Bin Size	1.0005
□ Use retention time tolerance		Min. Rel. Intensity	0.01	Min. Bin m/z	200.00
Ret. Time Tolerance	∞	Min. Ret. Time	10.00	Max. Bin m/z	1700.00

Knowledge base description

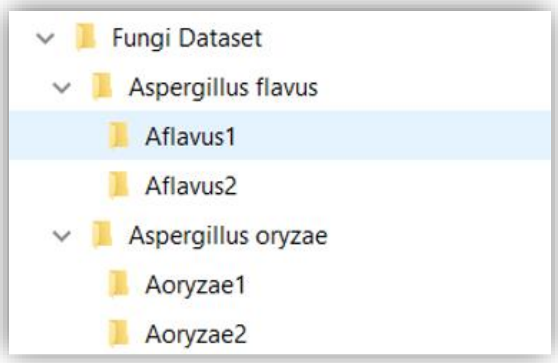
Step 1: Select a knowledge base

Step 2: Select a biological sample

Step 3: Select a directory with MS files

Figure 9 *The KB Manager interface.*

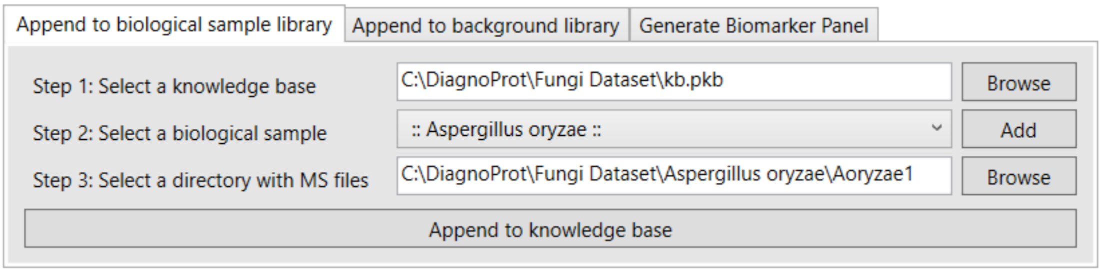
If the user has all the MS files (RAW, MS2, etc.) organized in a certain directory structure, the KB and all its biological conditions and clusters can be created with a single click. Figure 10 shows an example of a valid directory structure. All the biological conditions must have their own directory under the dataset root directory. Under each condition directory there must be a directory for each biological sample belonging to that condition. Finally, all MS files (including technical replicates) must under their respective biological sample directory.



- Fungi Dataset
  - Aspergillus flavus
    - Aflavus1
    - Aflavus2
  - Aspergillus oryzae
    - Aoryzae1
    - Aoryzae2

Figure 10 *Directory structure for automatic KB creation.*

Once a KB is created, the user can easily add new biological conditions to the KB and/or append new spectra to the conditions as soon as new data becomes available, as shown in Figure 11.


Append to biological sample library
Append to background library
Generate Biomarker Panel
Browse
Add
Browse
Append to knowledge base

Step 1: Select a knowledge base: C:\DiagnoProt\Fungi Dataset\kb.pkb

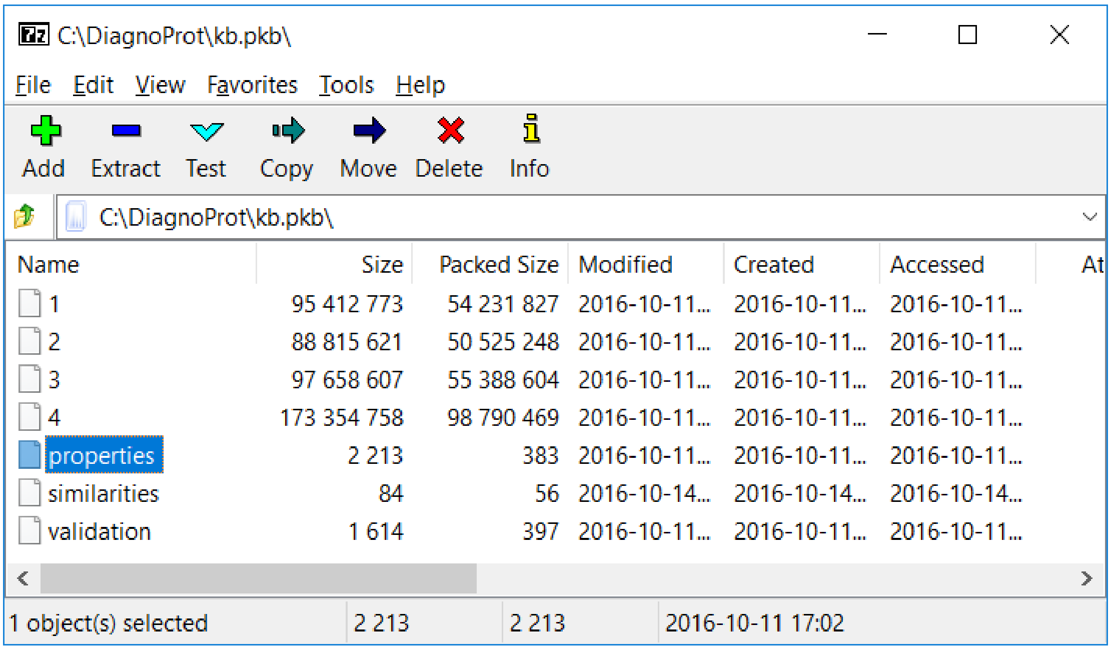
Step 2: Select a biological sample: :: Aspergillus oryzae ::

Step 3: Select a directory with MS files: C:\DiagnoProt\Fungi Dataset\Aspergillus oryzae\Aoryzae1

Figure 11 *Appending new data to the KB.*

5.3 The PKB file format

The KBs are stored in a PKB file, which is a compressed ZIP file containing serialized objects. Object serialization is a procedure that saves object-oriented data structures from computer memory into persistent media (e.g. hard-disk, etc.). The serialization is performed with Google Protocol Buffers, which is a very efficient serialization mechanism. This technology allows fast storage and retrieval of spectral clusters. In addition, since the KB is just a single compressed file, it can be easily shared with collaborators and end-users of the KB.



Name	Size	Packed Size	Modified	Created	Accessed	At
1	95 412 773	54 231 827	2016-10-11...	2016-10-11...	2016-10-11...	
2	88 815 621	50 525 248	2016-10-11...	2016-10-11...	2016-10-11...	
3	97 658 607	55 388 604	2016-10-11...	2016-10-11...	2016-10-11...	
4	173 354 758	98 790 469	2016-10-11...	2016-10-11...	2016-10-11...	
properties	2 213	383	2016-10-11...	2016-10-11...	2016-10-11...	
similarities	84	56	2016-10-14...	2016-10-14...	2016-10-14...	
validation	1 614	397	2016-10-11...	2016-10-11...	2016-10-11...	

1 object(s) selected 2 213 2 213 2016-10-11 17:02

Figure 12 *Example of a PKB file.*

Figure 12 shows an example of a PKB file opened with the software 7-Zip. Inside the file, you can see all the serialized objects. The numbered objects are the collections of spectral clusters, one for each biological condition in the KB. Since each condition has its own file, it is not necessary to load the entire KB into memory during computation, so this file format allows efficient use of computer memory. The properties object stores the parameters that are set during KB creation, the similarities object stores pre-computed similarities between biological conditions and the validation object stores information about the cross-validation procedure for classification.

5.4 Balance Score Distributions

Figure 13 shows the intensities distributions for estimated from spectra with different qualities and different precursor charge states from a *Mus musculus* dataset. The set of good spectra consist of spectra that were identified by Comet with very high Xcorr, while the bad spectra are the ones with the lowest Xcorr values. It is clear that 2+ intensities distributions vary greatly with the quality of the spectra, while 3+ distributions vary slightly. DiagnoProt uses these two default Balance distributions estimated from the two sets of good spectra, one for spectra with 2+ precursors and one for 3+ or more. Although this default distribution should be sufficient for typical applications, the user can easily provide other model distributions more suitable to their application.

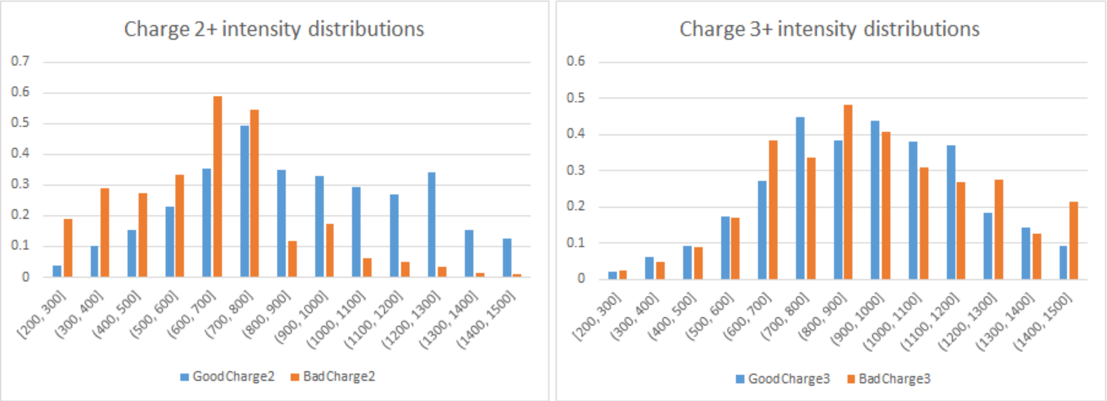
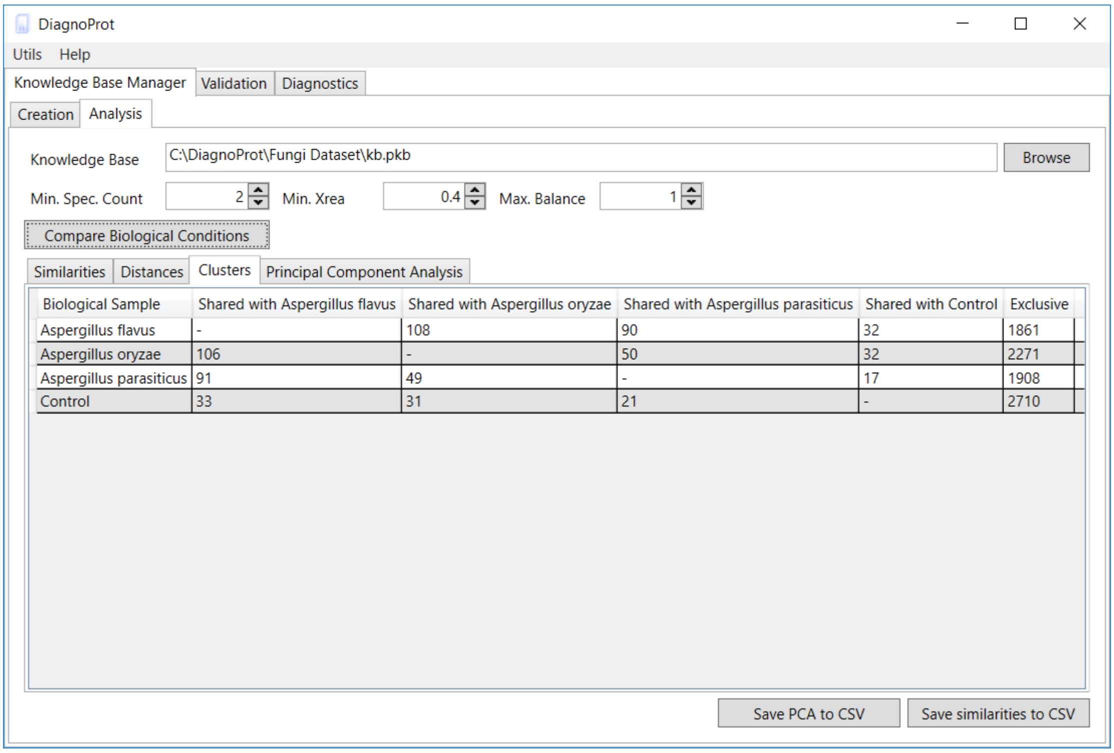


Figure 13 *Intensities distributions for 2+ and 3+ spectra.*

5.5 Discriminating Analysis



Knowledge Base: C:\DiagnoProt\Fungi Dataset\kb.pkb

Min. Spec. Count: 2 Min. Xrea: 0.4 Max. Balance: 1

Biological Sample	Shared with Aspergillus flavus	Shared with Aspergillus oryzae	Shared with Aspergillus parasiticus	Shared with Control	Exclusive
Aspergillus flavus	-	108	90	32	1861
Aspergillus oryzae	106	-	50	32	2271
Aspergillus parasiticus	91	49	-	17	1908
Control	33	31	21	-	2710

Figure 14 *KB Manager analysis interface.*

Figure 14 shows DiagnoProt's KB Manager / Analysis interface. When the search for discriminative clusters is completed, the Clusters tab shows a table with the numbers of clusters that are shared with each pair of biological condition, and the number of clusters that are exclusive of each condition.

It is interesting to note that the number of clusters of *A. flavus* condition shared with *A. oryzae* is 108, but the number of clusters shared between *A. oryzae* and *A. flavus* is 106. This happens because the previously defined relation "is similar to" is not transitive. Figure 15 shows an example of two cluster collections for conditions *A* and *B*. The numbered nodes are clusters, and an edge between two nodes means that the nodes are similar. Note that 3 is similar to 1, 1 is similar to 4, but 3 is not similar to 4. Therefore, when we count the number of clusters in A that occur in B we find 2, but when we count count the number of clusters in B that occur in A we find 3.

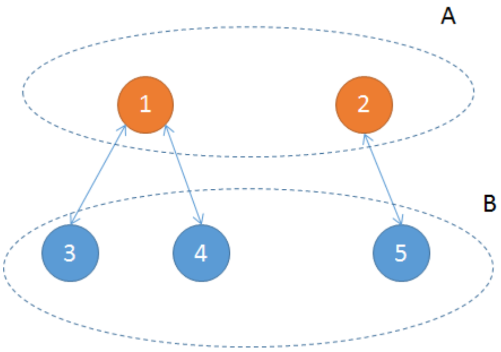
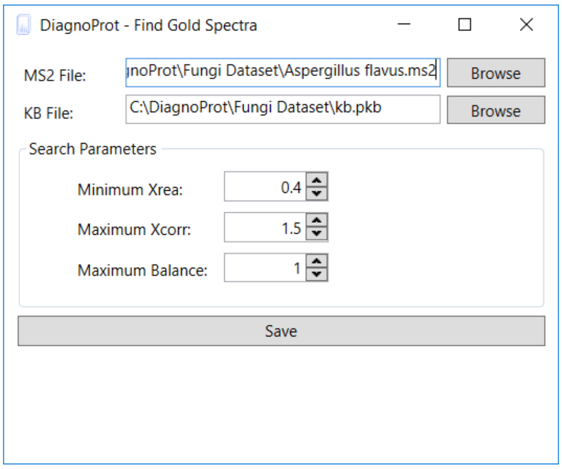


Figure 15 *Non-transitivity in cluster counting.*

If the user double-clicks a cell in the Clusters table shown in Figure 14, it will open the spectrum viewer in Figure 17. The viewer shows a plot of the binned spectra, and shows other relevant information for each spectrum: the MS file where it came from, the scan number of the original spectrum in its original file, the m/z of the precursor ion (CargedPrecursor), the charge state of the precursor ion, the Xrea and the Balance of the spectrum.

5.6 Shortlisting of new molecules



DiagnoProt - Find Gold Spectra

MS2 File: gnoProt\Fungi Dataset\Aspergillus flavus.ms2

KB File: C:\DiagnoProt\Fungi Dataset\kb.pkb

Search Parameters

Minimum Xrea: 0.4

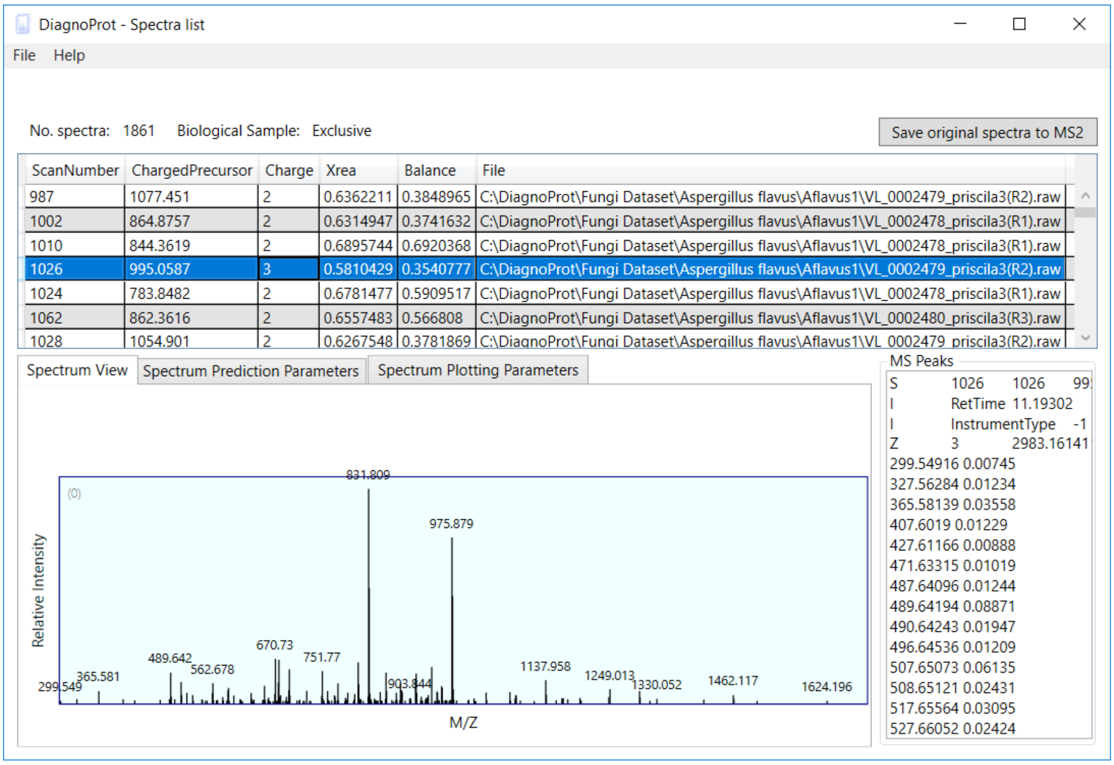
Maximum Xcorr: 1.5

Maximum Balance: 1

Figure 16 *Find Gold Spectra interface.*

The Find Gold Spectra tool shown in Figure 16 shortlists the high-quality discriminative spectra that could not be identified by PSM search. To find gold, the user must first save the original spectra using the spectrum viewer interface shown in Figure 17, and run Comet. Then, the Find Gold interface will ask for the file with the original spectra, the KB file, and the score thresholds: Min. Xrea, Max. Balance and Max. Xcorr. The Xrea and Balance thresholds are the same QC filters previously described, and the Max. Xcorr is the maximum allowed cross-correlation score for a spectrum to be considered unidentified. The default Max. Xcorr is 1.5, which is considered a very low Xcorr. Peptide assignments to spectra with such Xcorr score are considered unreliable, and are disregarded by post-processing tools (e.g. SEPro) for identification of reliable peptide assignments.

When Find Gold Spectra tool finishes the search, it shows all the high-quality unidentified spectra in the spectrum viewer (Figure 17), where the user can explore and save the spectra, and save all the original spectra (without binning transformation) in the viewer into the MS2 file for further analysis.



No. spectra: 1861 Biological Sample: Exclusive

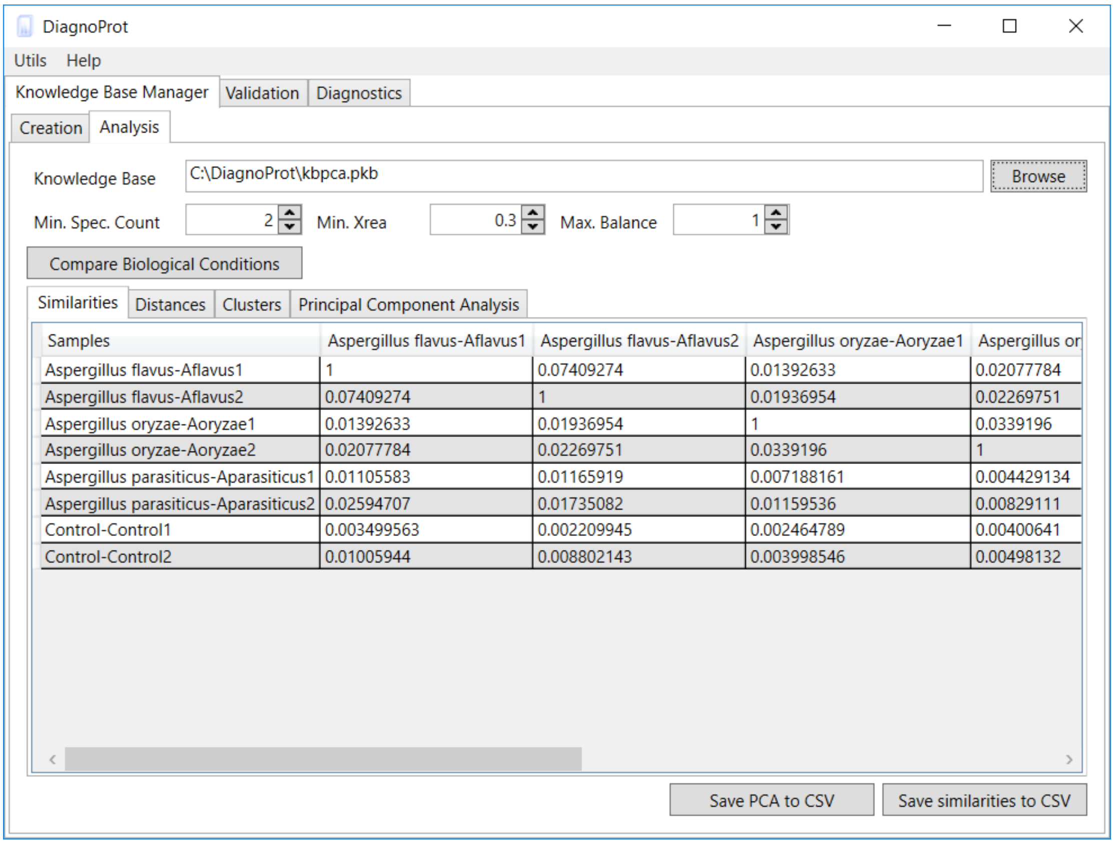
ScanNumber	ChargedPrecursor	Charge	Xrea	Balance	File
987	1077.451	2	0.6362211	0.3848965	C:\DiagnoProt\Fungi Dataset\Aspergillus flavus\Aflavus1\VL_0002479_priscila3(R2).raw
1002	864.8757	2	0.6314947	0.3741632	C:\DiagnoProt\Fungi Dataset\Aspergillus flavus\Aflavus1\VL_0002478_priscila3(R1).raw
1010	844.3619	2	0.6895744	0.6920368	C:\DiagnoProt\Fungi Dataset\Aspergillus flavus\Aflavus1\VL_0002478_priscila3(R1).raw
1026	995.0587	3	0.5810429	0.3540777	C:\DiagnoProt\Fungi Dataset\Aspergillus flavus\Aflavus1\VL_0002479_priscila3(R2).raw
1024	783.8482	2	0.6781477	0.5909517	C:\DiagnoProt\Fungi Dataset\Aspergillus flavus\Aflavus1\VL_0002478_priscila3(R1).raw
1062	862.3616	2	0.6557483	0.566808	C:\DiagnoProt\Fungi Dataset\Aspergillus flavus\Aflavus1\VL_0002480_priscila3(R3).raw
1028	1054.901	2	0.6267548	0.3781869	C:\DiagnoProt\Fungi Dataset\Aspergillus flavus\Aflavus1\VL_0002479_priscila3(R2).raw

```
S        1026     1026     99
I        RetTime 11.19302
I        InstrumentType   -1
Z        3        2983.16141
299.54916 0.00745
327.56284 0.01234
365.58139 0.03558
407.6019 0.01229
427.61166 0.00888
471.63315 0.01019
487.64096 0.01244
489.64194 0.08871
490.64243 0.01947
496.64536 0.01209
507.65073 0.06135
508.65121 0.02431
517.65564 0.03095
527.66052 0.02424
```

Figure 17 *Spectrum viewer.*

5.7 Principal Component Analysis

The KB Manager Analysis tab has a tool that graphically represents biological samples that are closely related in a two-dimensional plot. When the Compare Biological Conditions button is clicked, in addition the previously described search for discriminative clusters, it also compares how biological samples are related to each other by computing the Jaccard index. The Similarities tab (Figure 18) shows the Jaccard indexes between all pairs of biological samples.



Knowledge Base C:\DiagnoProt\kbpca.pkb

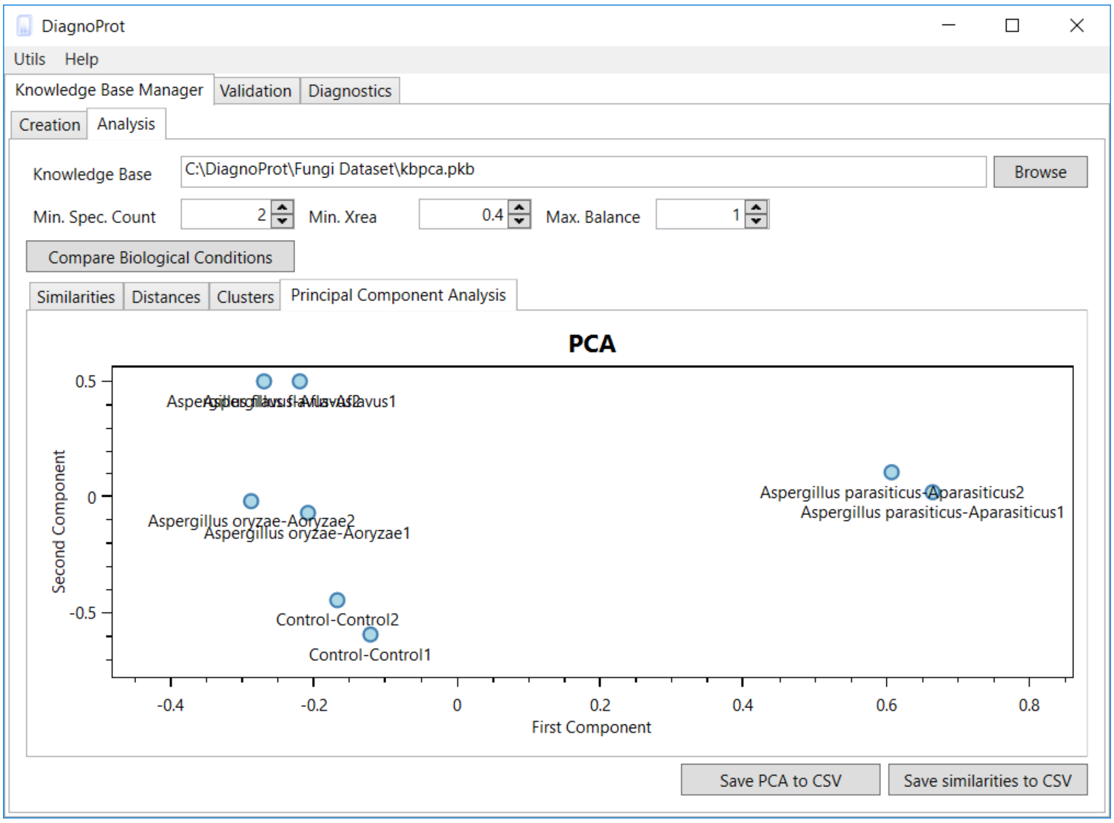
Min. Spec. Count 2 Min. Xrea 0.3 Max. Balance 1

Samples	Aspergillus flavus-Aflavus1	Aspergillus flavus-Aflavus2	Aspergillus oryzae-Aoryzae1	Aspergillus or
Aspergillus flavus-Aflavus1	1	0.07409274	0.01392633	0.02077784
Aspergillus flavus-Aflavus2	0.07409274	1	0.01936954	0.02269751
Aspergillus oryzae-Aoryzae1	0.01392633	0.01936954	1	0.0339196
Aspergillus oryzae-Aoryzae2	0.02077784	0.02269751	0.0339196	1
Aspergillus parasiticus-Aparasiticus1	0.01105583	0.01165919	0.007188161	0.004429134
Aspergillus parasiticus-Aparasiticus2	0.02594707	0.01735082	0.01159536	0.00829111
Control-Control1	0.003499563	0.002209945	0.002464789	0.00400641
Control-Control2	0.01005944	0.008802143	0.003998546	0.00498132

Figure 18 *Similarities between biological samples.*

The distance between two biological samples with spectral cluster collections A and B is defined as $1 - J(A, B)$. The Distances tab in Figure 1 shows the distances between all biological samples in the KB. Let d_{ij} be the distance between sample i and j. Using this matrix of distances, we can represent the relation of a biological condition with all others as a vector of distances $\overrightarrow{d_\iota} = (d_{i1}, d_{i2}, \cdots, d_{ij} \cdots, d_{in})$, where n is the number of biological samples in the KB. Then a Principal Component Analysis (PCA) is performed in this set of vectors $\overrightarrow{d_1}, \cdots, \overrightarrow{d_n}$ to reduce its dimension from n to 2 by making a projection into the directions of highest variance in the data. The projected vectors are then plotted in two-dimensional chart shown in Figure 19.

With this procedure, it is possible to graphically represent distances between biological samples in the KB. It is expected that biological samples belonging to the same biological condition will share much biological material, and therefore they may be graphically closer in the PCA plot.

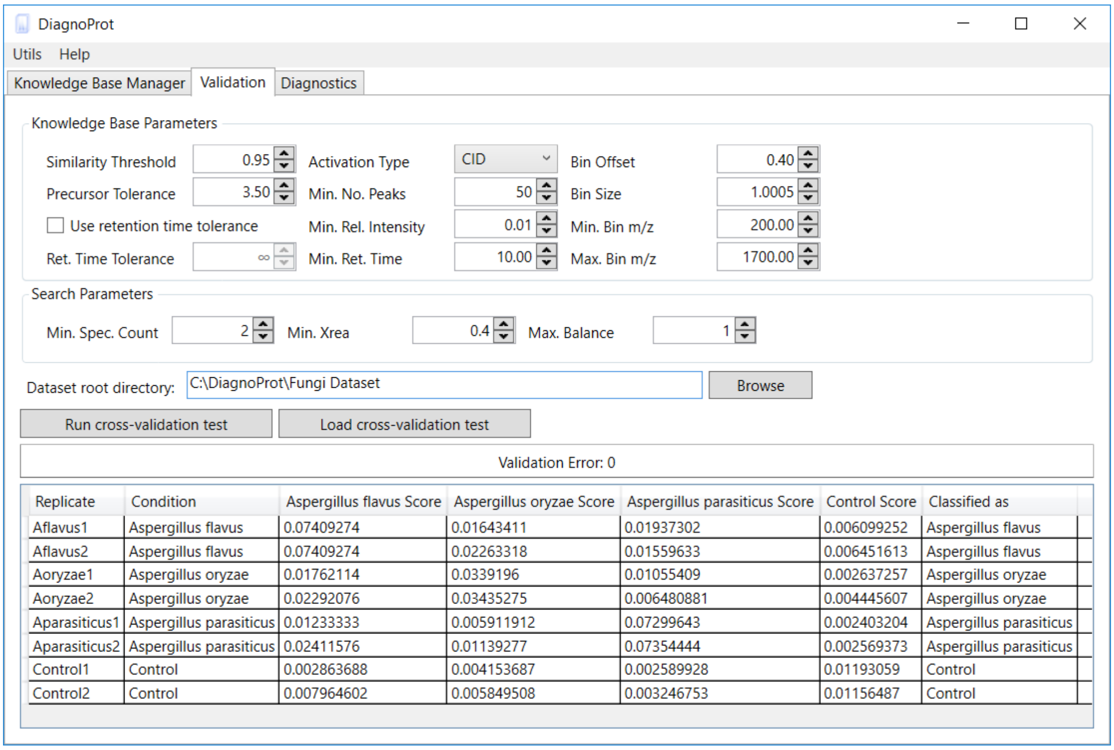


Knowledge Base C:\DiagnoProt\Fungi Dataset\kbpca.pkb

Min. Spec. Count 2 Min. Xrea 0.4 Max. Balance 1

Figure 19 *PCA plot.*

5.8 Cross-validation



Knowledge Base Parameters

Similarity Threshold 0.95 Activation Type CID Bin Offset 0.40

Precursor Tolerance 3.50 Min. No. Peaks 50 Bin Size 1.0005

□ Use retention time tolerance Min. Rel. Intensity 0.01 Min. Bin m/z 200.00

Ret. Time Tolerance ∞ Min. Ret. Time 10.00 Max. Bin m/z 1700.00

Search Parameters

Min. Spec. Count 2 Min. Xrea 0.4 Max. Balance 1

Dataset root directory: C:\DiagnoProt\Fungi Dataset

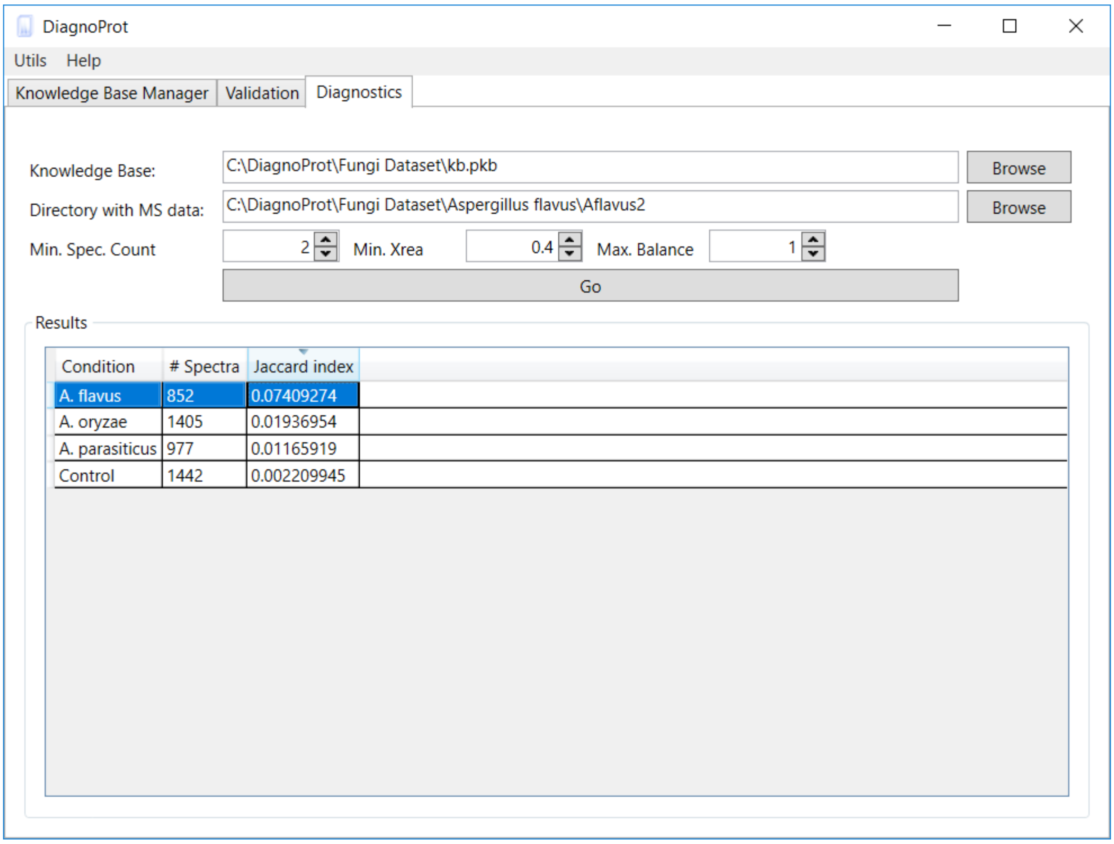
Validation Error: 0

Replicate	Condition	Aspergillus flavus Score	Aspergillus oryzae Score	Aspergillus parasiticus Score	Control Score	Classified as
Aflavus1	Aspergillus flavus	0.07409274	0.01643411	0.01937302	0.006099252	Aspergillus flavus
Aflavus2	Aspergillus flavus	0.07409274	0.02263318	0.01559633	0.006451613	Aspergillus flavus
Aoryzae1	Aspergillus oryzae	0.01762114	0.0339196	0.01055409	0.002637257	Aspergillus oryzae
Aoryzae2	Aspergillus oryzae	0.02292076	0.03435275	0.006480881	0.004445607	Aspergillus oryzae
Aparasiticus1	Aspergillus parasiticus	0.01233333	0.005911912	0.07299643	0.002403204	Aspergillus parasiticus
Aparasiticus2	Aspergillus parasiticus	0.02411576	0.01139277	0.07354444	0.002569373	Aspergillus parasiticus
Control1	Control	0.002863688	0.004153687	0.002589928	0.01193059	Control
Control2	Control	0.007964602	0.005849508	0.003246753	0.01156487	Control

Figure 20 *Cross-validation interface.*

Figure 20 shows DiagnoProt's cross-validation interface. The KB creation and the QC search parameters must be set. Also, all MS files must be organized in the directory structure for automatic KB creation defined in section 5.2. The user can stop it at any time, and the LOOCV will continue from where it left when the user starts the validation again. When LOOCV is over, the validation error is shown as well as the individual classifier assignments for all biological samples.

5.9 Classification of unknown profiles



Knowledge Base: C:\DiagnoProt\Fungi Dataset\kb.pkb

Directory with MS data: C:\DiagnoProt\Fungi Dataset\Aspergillus flavus\Aflavus2

Min. Spec. Count 2 Min. Xrea 0.4 Max. Balance 1

Condition	# Spectra	Jaccard index
A. flavus	852	0.07409274
A. oryzae	1405	0.01936954
A. parasiticus	977	0.01165919
Control	1442	0.002209945

Figure 21 *Diagnostics interface.*

If after validation the user is confident that the classifier and the KB will generalize well out-of-sample, then he can use the spectrum profile classifier in the Diagnostics tab (Figure 21) for unknown samples. The interface asks for a KB, a directory with MS files from the unknown samples, and QC filter thresholds. Then, the classifier shows a table with the Jaccard indexes between the unknown sample and the KB conditions as well as the number of unknown spectra that occur in each condition.

5.10 Performance in datasets

The methodology was applied to a dataset of three species of fungi of the *Aspergillus* genus. Using the default parameters form KB creation previously described, a

KB was created and a search for high-quality exclusive spectra for each condition was performed. The exclusive spectra were then analyzed by the Comet search engine. Finally, using the Comet results, a search for gold spectra was performed. The results are summarized in Table 1. DiagnoProt found all these high-quality spectra that are exclusive occurring in their respective conditions. Most of them could not be identified by Comet, nonetheless DiagnoProt could shortlist them as potential new molecules with discriminative power among those fungi species.

Biological Condition	*Exclusive Clusters*	*Gold Clusters*
A. flavus	1861	1457
A. oryzae	2272	1289
A. parasiticus	1908	1620

Table 1 *Discriminative spectra of Aspergillus dataset.*

Table 2 shows the cross-validation results. All the biological replicates were classified correctly by the spectral profile classifier. Even the control samples were classified correctly.

		Scores				
Sample	*Condition*	*A. flavus*	*A. oryzae*	*A. parasiticus*	*Control*	*Assignment*
Aflavus1	A. flavus	**0.074**	0.016	0.019	0.006	A. flavus
Aflavus2	A. flavus	**0.074**	0.023	0.016	0.006	A. flavus
Aoryzae1	A. oryzae	0.018	**0.034**	0.011	0.003	A. oryzae
Aoryzae2	A. oryzae	0.023	**0.034**	0.006	0.004	A. oryzae
Aparasiticus1	A. parasiticus	0.012	0.006	**0.073**	0.002	A. parasiticus
Aparasiticus2	A. parasiticus	0.024	0.011	**0.074**	0.003	A. parasiticus
Control1	Control	0.003	0.004	0.003	**0.012**	Control
Control2	Control	0.008	0.006	0.003	**0.012**	Control

Table 2 *LOOCV results of Aspergillus dataset.*

A PCA plot of the distances between the biological replicates was obtained and is show in Figure 22. Note that the samples belonging to the same condition are naturally clustered together in the chart.

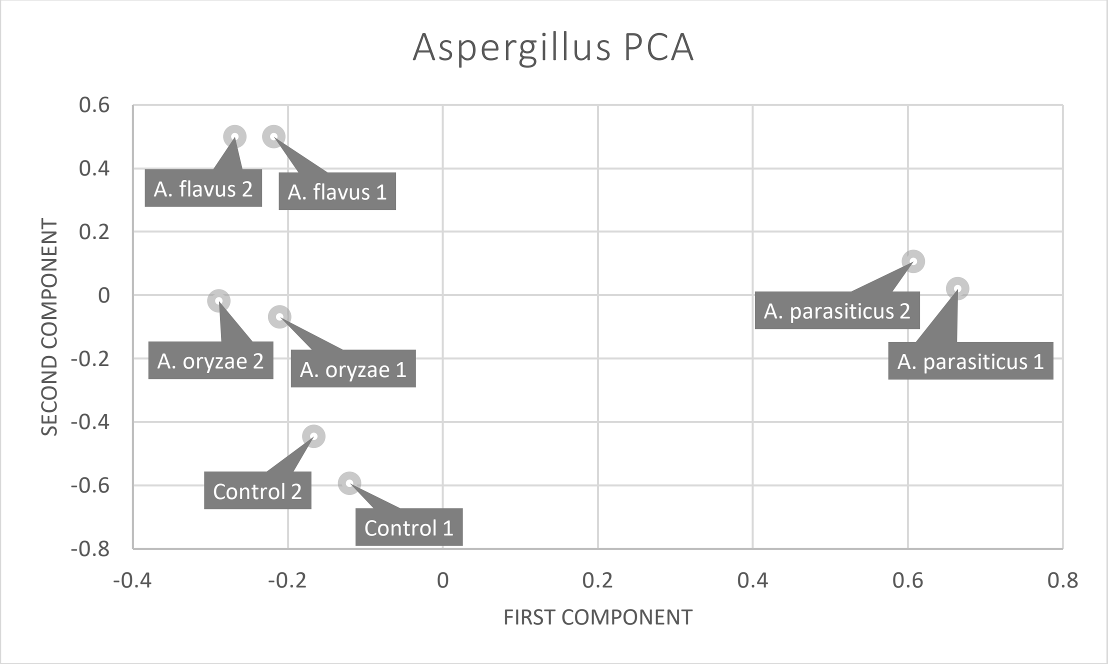


Figure 22 *PCA plot of Aspergillus dataset.*

DiagnoProt was also applied to a dataset of chronical kidney diseases with data from 12 patients diagnosed with lupus nephropathy (LN) and 11 patients with IgA nephropathy (IgAN).

The PCA plot in Figure 23 shows that LN and IgAN samples are naturally clustered together, and Table 3 shows that all samples were correctly classified according to LOOCV.

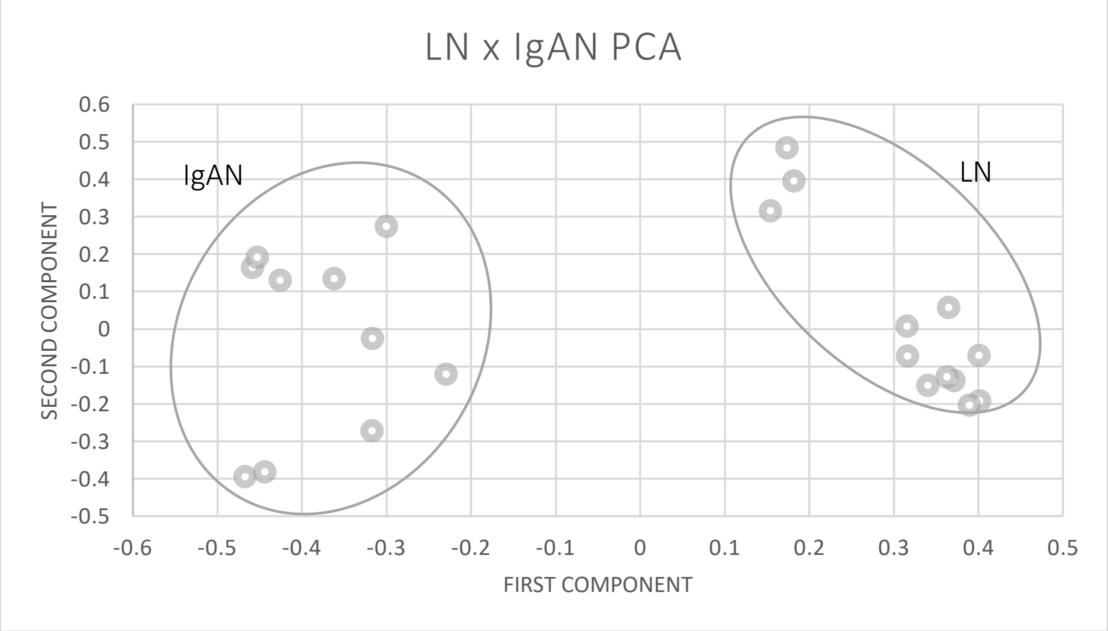


Figure 23 *PCA plot of kidney diseases dataset.*

		Scores		
Sample	*Condition*	*IgAN score*	*LN score*	*Assignment*
O.27	IgAN	**0.046**	0.024	IgAN
O.29	IgAN	**0.037**	0.024	IgAN
O.44	IgAN	**0.036**	0.017	IgAN
O.53	IgAN	**0.050**	0.020	IgAN
O.54	IgAN	**0.037**	0.020	IgAN
O.60	IgAN	**0.056**	0.024	IgAN
O.61	IgAN	**0.060**	0.025	IgAN
O.64	IgAN	**0.051**	0.023	IgAN
O.78	IgAN	**0.065**	0.030	IgAN
O.81	IgAN	**0.061**	0.029	IgAN
O.97	IgAN	**0.063**	0.027	IgAN
O.15	LN	0.004	**0.030**	LN
O.18	LN	0.004	**0.038**	LN
O.25	LN	0.004	**0.039**	LN
O.28	LN	0.003	**0.022**	LN
O.30	LN	0.004	**0.040**	LN
O.36	LN	0.006	**0.016**	LN
O.41	LN	0.003	**0.025**	LN
O.42	LN	0.005	**0.043**	LN
O.49	LN	0.006	**0.025**	LN
O.51	LN	0.004	**0.030**	LN
O.56	LN	0.004	**0.033**	LN
O.8_6	LN	0.003	**0.017**	LN

Table 3 *LOOCV results of kidney diseases dataset.*

6 Discussion

We have successfully developed a methodology and a software that can discover new unidentified molecules that have discriminative power among biological conditions, and therefore, the knowledge of these new molecules can lead to better understanding of biological systems as well as the construction of better diagnosis methods and tools. We foresee that this new information will be very useful for creation of new diagnosis methods that will be able to discriminate among disease states based on tandem spectral profiling of less invasive samples, such as urine or saliva. The development of diagnosis systems based on this methodology has a huge potential to revolutionize medical technology. Furthermore, it is important to emphasize that our methodology is based on spectral profiling, not just based on a single protein fingerprint as existing commercial approaches (e.g., Biotyper). Thus, any biological sample that could be analyzed by mass spectrometry (e.g., lipids, peptides, proteins, etc.) can also be analyzed by discriminating spectral cluster analysis. Since our method works by pair-wise comparisons of spectral similarity and pattern matching, it is indeed a general tool for discovery of discriminative biological factors and classification of spectral profiles by mass spectrometry. We are convinced that this approach is a step forward in terms of personalized diagnosis as the protein profile from a patient will be considered; this test could ultimately even be deployed as a public health service assisting in diagnosis of special cases when existing approaches are elusive.

7 Conclusion

We have successfully developed and implemented a methodology for shortlisting discriminative spectra among biological conditions that PSM search could not identify. For that, DiagnoProt communicate with Comet search engine. The knowledge base data structure allows fast access and efficient storage of spectra. DiagnoProt can also graphically represent distances among biological conditions and biological samples base on spectral profiling comparisons. Spectral profile classification could classify different fungi samples, and showed promising results in the kidney disease dataset.

This work resulted in a patent filed in the USA (Anexo I) and an application note submitted to Bioinformatics (Anexo II).

DiagnoProt is available at http://patternlabforproteomics.org/diagnoprot/.

8 Bibliography

[1] ABU-MOSTAFA, Y. S.; MAGDON-ISMAIL, M.; LIN, H.-T. **Learning From Data**. AMLBook, 2012.

[2] AQUINO, P. F.; FISCHER, J. S. G.; NEVES-FERREIRA, A. G. C.; et al. Are gastric cancer resection margin proteomic profiles more similar to those from controls or tumors? **Journal of proteome research**, v. 11, n. 12, p. 5836–5842. doi: 10.1021/pr300612x, 2012.

[3] BORGES, D.; PEREZ-RIVEROL, Y.; NOGUEIRA, F. C. S.; et al. Effectively addressing complex proteomic search spaces with peptide spectrum matching. **Bioinformatics (Oxford, England)**, v. 29, n. 10, p. 1343–1344. doi: 10.1093/bioinformatics/btt106, 2013.

[4] CARVALHO, P. C.; BARBOSA, V. C. **Um Ambiente Computacional para Proteômica**. Doctorate Thesis, Rio de Janeiro: COPPE/UFRJ. Retrieved from http://www3.cos.ufrj.br/index.php?option=com_publicacao&task=visualizar&id=2119, 2010, March 31.

[5] CARVALHO, P. C.; FISCHER, J. S. G.; XU, T.; et al. Search engine processor: Filtering and organizing peptide spectrum matches. **Proteomics**, v. 12, n. 7, p. 944–949. doi: 10.1002/pmic.201100529, 2012.

[6] CARVALHO, P. C.; LIMA, D. B.; LEPREVOST, F. V.; et al. Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0. **Nature Protocols**, v. 11, n. 1, p. 102–117. doi: 10.1038/nprot.2015.133, 2015.

[7] ENG, J. K.; HOOPMANN, M. R.; JAHAN, T. A.; et al. A deeper look into Comet--implementation and features. **Journal of the American Society for Mass Spectrometry**, v. 26, n. 11, p. 1865–1874. doi: 10.1007/s13361-015-1179-x, 2015.

[8] ENG, J. K.; JAHAN, T. A.; HOOPMANN, M. R. Comet: an open-source MS/MS sequence database search tool. **Proteomics**, v. 13, n. 1, p. 22–24. doi: 10.1002/pmic.201200439, 2013.

[9] ENG, J. K.; MCCORMACK, A. L.; YATES, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. **Journal of the American Society for Mass Spectrometry**, v. 5, n. 11, p. 976–989. doi: 10.1016/1044-0305(94)80016-2, 1994.

[10] FENN, J. B.; MANN, M.; MENG, C. K.; WONG, S. F.; WHITEHOUSE, C. M. Electrospray ionization for mass spectrometry of large biomolecules. **Science**, v. 246, n. 4926, p. 64–71. doi: 10.1126/science.2675315, 1989.

[11] GRISS, J.; PEREZ-RIVEROL, Y.; LEWIS, S.; et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. **Nature Methods**, v. 13, n. 8, p. 651–656. doi: 10.1038/nmeth.3902, 2016.

[12] JAMES, P. Protein identification in the post-genome era: the rapid rise of proteomics. **Quarterly Reviews of Biophysics**, v. 30, n. 4, p. 279–331, 1997.

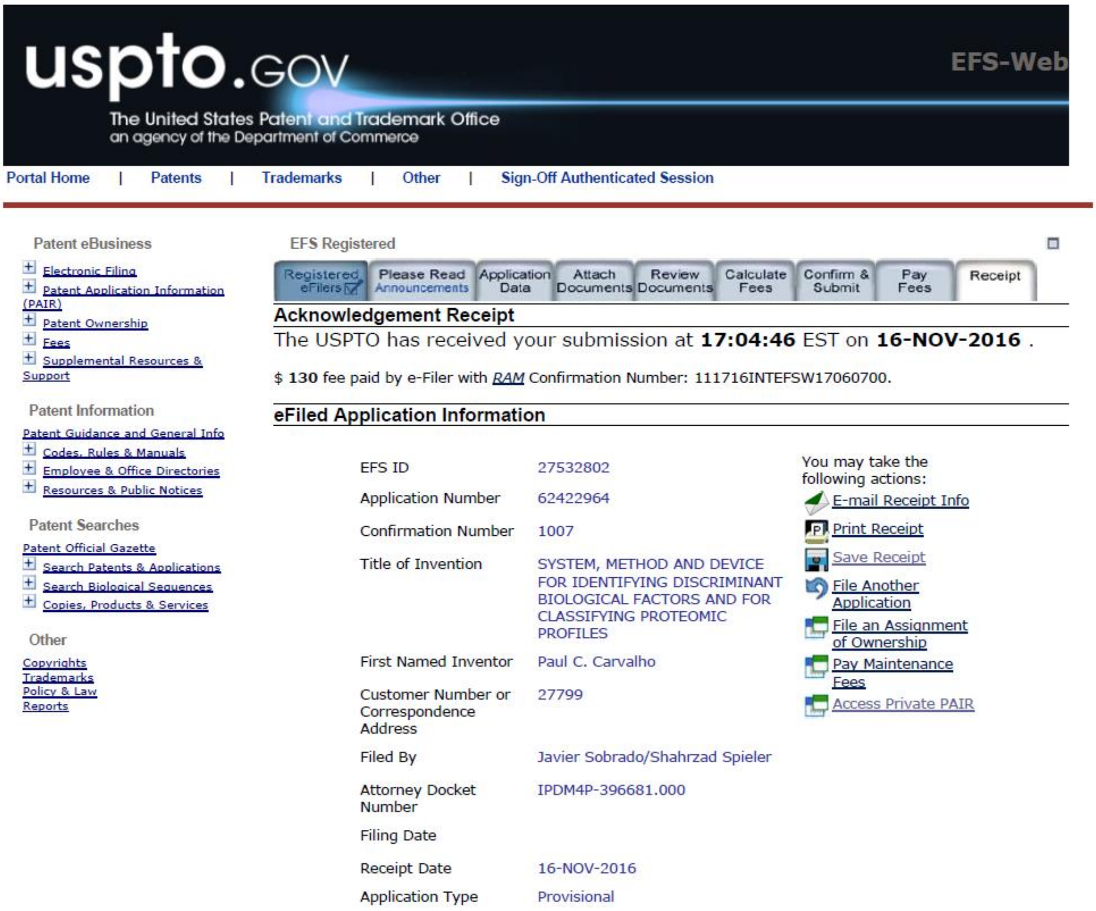
[13] LIMA, DIOGO BORGES; FRANÇA, FELIPE MAIA GALVÃO; CARVALHO, PAULO COSTA. **Método computacional para identificação**

de peptídeos marcados com fenil-isotiocianato e analisados por cromatografia líquida acoplada a espectrometria de massa em tandem. Rio de Janeiro: Federal University of Rio de Janeiro, 2013, February.

[14] NA, S.; PAEK, E. Quality Assessment of Tandem Mass Spectra Based on Cumulative Intensity Normalization. **Journal of Proteome Research**, v. 5, n. 12, p. 3241–3248. doi: 10.1021/pr0603248, 2006.

[15] WELKER, M. Proteomics for routine identification of microorganisms. **PROTEOMICS**, v. 11, n. 15, p. 3143–3153. doi: 10.1002/pmic.201100049, 2011.

Anexo I


uspto.GOV
The United States Patent and Trademark Office
an agency of the Department of Commerce
EFS-Web
Portal Home | Patents | Trademarks | Other | Sign-Off Authenticated Session
Patent eBusiness
Electronic Filing
Patent Application Information (PAIR)
Patent Ownership
Fees
Supplemental Resources & Support
Patent Information
Patent Guidance and General Info
Codes, Rules & Manuals
Employee & Office Directories
Resources & Public Notices
Patent Searches
Patent Official Gazette
Search Patents & Applications
Search Biological Sequences
Copies, Products & Services
Other
Copyrights
Trademarks
Policy & Law
Reports
EFS Registered
Registered eFilers
Please Read Announcements
Application Data
Attach Documents
Review Documents
Calculate Fees
Confirm & Submit
Pay Fees
Receipt

Acknowledgement Receipt

The USPTO has received your submission at **17:04:46** EST on **16-NOV-2016** .

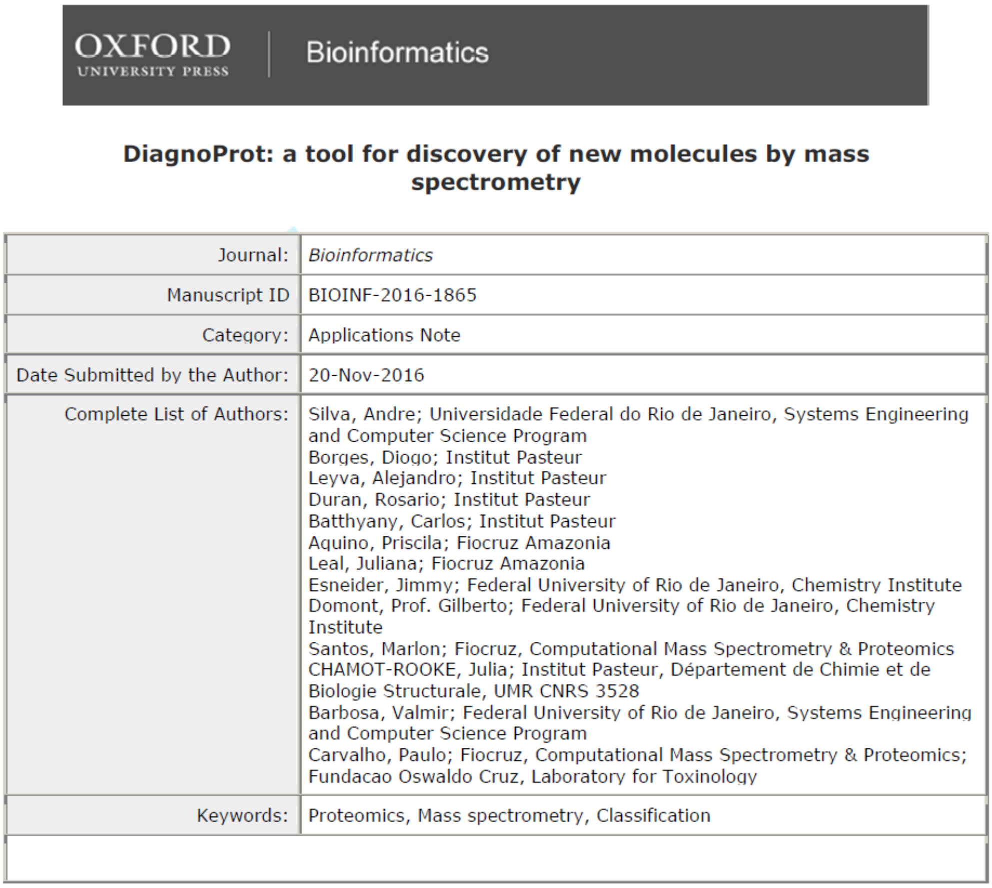
$ **130** fee paid by e-Filer with *RAM* Confirmation Number: 111716INTEFSW17060700.

eFiled Application Information

EFS ID	27532802
Application Number	62422964
Confirmation Number	1007
Title of Invention	SYSTEM, METHOD AND DEVICE FOR IDENTIFYING DISCRIMINANT BIOLOGICAL FACTORS AND FOR CLASSIFYING PROTEOMIC PROFILES
First Named Inventor	Paul C. Carvalho
Customer Number or Correspondence Address	27799
Filed By	Javier Sobrado/Shahrzad Spieler
Attorney Docket Number	IPDM4P-396681.000
Filing Date	
Receipt Date	16-NOV-2016
Application Type	Provisional

You may take the following actions:

Anexo II



OXFORD UNIVERSITY PRESS | Bioinformatics

DiagnoProt: a tool for discovery of new molecules by mass spectrometry

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2016-1865
Category:	Applications Note
Date Submitted by the Author:	20-Nov-2016
Complete List of Authors:	Silva, Andre; Universidade Federal do Rio de Janeiro, Systems Engineering and Computer Science Program Borges, Diogo; Institut Pasteur Leyva, Alejandro; Institut Pasteur Duran, Rosario; Institut Pasteur Batthyany, Carlos; Institut Pasteur Aquino, Priscila; Fiocruz Amazonia Leal, Juliana; Fiocruz Amazonia Esneider, Jimmy; Federal University of Rio de Janeiro, Chemistry Institute Domont, Prof. Gilberto; Federal University of Rio de Janeiro, Chemistry Institute Santos, Marlon; Fiocruz, Computational Mass Spectrometry & Proteomics CHAMOT-ROOKE, Julia; Institut Pasteur, Département de Chimie et de Biologie Structurale, UMR CNRS 3528 Barbosa, Valmir; Federal University of Rio de Janeiro, Systems Engineering and Computer Science Program Carvalho, Paulo; Fiocruz, Computational Mass Spectrometry & Proteomics; Fundacao Oswaldo Cruz, Laboratory for Toxinology
Keywords:	Proteomics, Mass spectrometry, Classification