MÉTODO COMPUTACIONAL PARA IDENTIFICAÇÃO DE PEPTÍDEOS MARCADOS COM FENIL-ISOTIOCIANATO E ANALISADOS POR CROMATOGRAFIA LÍQUIDA ACOPLADA A ESPECTROMETRIA DE MASSA EM TANDEM

Diogo Borges Lima

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: Felipe Maia Galvão França
Paulo Costa Carvalho

Rio de Janeiro
Fevereiro de 2013

MÉTODO COMPUTACIONAL PARA IDENTIFICAÇÃO PEPTÍDEOS MARCADOS COM FENIL-ISOTIOCIANATO E ANALISADOS POR CROMATOGRAFIA LÍQUIDA ACOPLADA A ESPECTROMETRIA DE MASSA EM TANDEM

Diogo Borges Lima

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. Felipe Maia Galvão França, Ph.D.

_____
Dr. Paulo Costa Carvalho, D.Sc.

_____
Prof. Valmir Carneiro Barbosa, Ph.D.

_____
Dr. Tiago Santana Balbuena, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
FEVEREIRO DE 2013.

Lima, Diogo Borges

Método computacional para identificação de peptídeos marcados com fenil-isotiocianato e analisados por cromatografia líquida acoplada a espectrometria de massa em tandem / Diogo Borges Lima. – Rio de Janeiro: UFRJ/COPPE, 2013.

XI, 50 p.: il.; 29,7 cm.

Orientadores: Felipe Maia Galvão França

Paulo Costa Carvalho

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2013.

Referências Bibliográficas: p. 45-47.

1. Proteômica Computacional. 2. Reconhecimento de Padrões. 3. Espectrometria de Massa. I. França, Felipe Maia Galvão; Carvalho, Paulo Costa. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Dedico esta dissertação à minha mãe e aos meus avós,

Mônica Maria Borges Lima,
Rita do Amaral Borges e
Benedito de Aragão Borges

Agradecimentos

Agradeço, em primeiro lugar, a Deus pelas oportunidades que me foram dadas e às pessoas que conheci, as quais me proporcionaram a evolução do aprendizado.

Meus agradecimentos também à minha mãe, Mônica Maria Borges Lima, e aos meus avós, Rita do Amaral Borges e Benedito de Aragão Borges por sempre terem me apoiado ao longo de minha vida, e por terem me educado, para que hoje eu pudesse ser a pessoa que sou. Não posso deixar de agradecer à minha irmã, Aline Thaís Borges Lima, por ter me dado forças nessa jornada do Mestrado.

Agradeço ao Dr. Paulo Costa Carvalho, pois além de ser um orientador, é um grande amigo que embarcou junto comigo nesse desafio multidisciplinar; ao Felipe da Veiga Leprevost por fazer parte do grupo de proteômica computacional o qual faço parte, e me ajudar em vários momentos no desenvolvimento da dissertação; e ao orientador Felipe Maia Galvão França pelo apoio na realização do curso de Mestrado.

E por fim, agradeço também ao CNPq pela ajuda financeira, para que esta dissertação pudesse ser realizada.

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

MÉTODO COMPUTACIONAL PARA IDENTIFICAÇÃO DE PEPTÍDEOS MARCADOS COM FENIL-ISOTIOCIANATO E ANALISADOS POR CROMATOGRAFIA LÍQUIDA ACOPLADA A ESPECTROMETRIA DE MASSA EM TANDEM

Diogo Borges Lima

Fevereiro/2013

Orientadores: Felipe Maia Galvão França
Paulo Costa Carvalho

Programa: Engenharia de Sistemas e Computação

A proteômica é uma ciência que faz uso da inteligência artificial para realizar a identificação, quantificação e caracterização de modificações pós-traducionais que podem ocorrer em proteínas nos diversos organismos. Este trabalho apresenta uma nova metodologia computacional e experimental capaz de aumentar consideravelmente a sensibilidade na identificação de peptídeos por cromatografia líquida e espectrometria de massas. Para isso, apresentamos uma ferramenta, denominada *Spectrum Identification Machine* (SIM), na qual implementamos esta metodologia confrontando espectros teóricos, gerados a partir de um banco de dados de sequências proteicas, com espectros experimentais. O aumento da sensibilidade é obtido através de uma marcação química nos peptídeos, denominada fenil-isotiocianato (PITC), que intensifica o íon *b1*, fazendo com que ele seja o mais intenso do espectro em uma determinada região. Criamos uma lógica capaz de explorar essa informação, e a programamos no SIM, fazendo com que a complexidade do espaço de busca diminua e, consequentemente, aumente a sensibilidade.

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

COMPUTATIONAL METHOD FOR IDENTIFYING PEPTIDES LABELED WITH PHENYLISOTHIOCYANATE AND ANALISED BY LIQUID CHROMATOGRAPHY COUPLED TO TANDEM MASS SPECTROMETRY

Diogo Borges Lima

February /2013

Advisors: Felipe Maia Galvão França
Paulo Costa Carvalho

Department: Systems Engineering and Computer Science

Proteomics is a science that heavily relies on artificial intelligence and pattern recognition to identify, quantitate, and characterize the various forms of proteins (e.g., post-translational modifications) from biological systems. This thesis presents a new computational and experimental method to effectively identify peptides coupled in solution with *Phenylisothiocyanate* (PITC) and analysed by tandem mass spectrometry. We present our strategy as a tool entitled *Spectrum Identification Machine* (SIM). SIM stands out from existing tools as it is significantly more sensitive (~100%); this is achieved by capitalizing on the high intensity of the b1-type ion of PITC peptides; this allows to sequence the first amino acid to reduce the search space to be queried by comparing theoretical spectra with experimental ones.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas, símbolos e unidades

CID	Dissociação por Colisão Induzida; do inglês, *Collision-Induced Dissociation*
Da	Dalton, unidade de massa atômica
ESI	Ionização por eletro spray; do inglês, *Electrospray Ionization*
ETD	Dissociação por Transferência de Elétron; do inglês, *Electron-Transfer Dissociation*
FDR	Taxa de Falsos Positivos; do inglês, *False Discovery Rate*
HCD	Dissociação por maior energia de colisão; do inglês, *Higher-Energy Collisional Dissociation*
HPP	Projeto Proteôma Humano; do inglês, *Human Proteome Project*
K	Aminoácido Lisina
LC/LC ou LC 2D	Cromatografia Líquida bi-dimensional
LTQ	*Liner Trap Quadrupolo*
MH	Massa isotópica acrescida do cátion Hidrogênio
MS1	Espectro do perfil de massas
MS2 ou MS/MS	Espectro do perfil de massas de íons dissociados
m/z	Razão massa/carga
MALDI	Ionização por dessorção a *laser* assistida por matriz; do inglês, *Matrix-assisted laser desorption/ionization*
MudPIT	*Multidimensional Protein Identification Technology*
PITC	Fenil-isotiocianato; do inglês, *Phenylisothiocyanate*
PTM	Modificação pós-traducional; do inglês, *Post-translational Modification*
ppm	Parte por milhão
R	Aminoácido Arginina
RP	Fase reversa; do inglês, *Reverse Phase*
SIM	*Spectrum Identification Machine*

1 Introdução

A Proteômica é uma ciência multidisciplinar onde cada vez mais a Ciência da Computação desempenha um papel fundamental. Nascida da Bioquímica, esta ciência hoje possibilita estudar amostras biológicas complexas (e.g., fluidos biológicos, lisados celulares), permitindo a identificação e quantificação de milhares de proteínas com a utilização de espectrômetros de massa de alta resolução. Sendo assim, estudos médicos, bioquímicos e biotecnológicos utilizam a proteômica para estudar patologias, sistemas biológicos e desenvolvimento de novas tecnologias.

1.1 Um breve histórico

No início do Projeto Genoma Humano, datado no final do século XX, acreditava-se que existiam cerca de 100 mil tipos de proteínas nos seres humanos. Entretanto, hoje sabe-se que este número fica em torno de 20 mil, dos quais grande parte já teriam suas funções conhecidas[3][21]. A proteína desempenha um papel fundamental nos organismos biológicos.

O termo "Proteômica" foi primeiramente adotado em 1997[26], fazendo uma analogia com o termo genômica, que é o estudo em larga escala dos genes. "Proteômica Computacional" é utilizada para fazer referência a análises experimentais de proteínas em grande escala. A palavra "Proteôma" é a aglutinação dos termos proteína e genoma, e foi cunhada por Marc Wilkins em 1994, enquanto desenvolvia sua tese de Doutorado[20]. O proteôma compreende todas as formas de proteínas, incluindo as modificações pós-traducionais (PTM's). Seria o estudo completo das proteínas que foram produzidas por um determinado genoma.

Criado em setembro de 2000, o Projeto Proteôma Humano tem como objetivo identificar e caracterizar, no mínimo, um produto proteico a partir de cada um dos 20.300 genes codificadores de proteínas (Human Proteome Project (HPP), 2010).

A proteômica *shotgun* é uma técnica do tipo *bottom-up*[1] que objetiva identificar milhares de proteínas em misturas complexas através da digestão das mesmas, utilizando a cromatografia líquida acoplada à espectrometria de massa.

[1] Proteômica *bottom-up* é um método para identificar as sequências peptídicas de proteínas e suas modificações pós-traducionais utilizando a digestão enzimática antes da análise por espectrometria de massa.[15]

É praticamente impérvio estudar algum tipo de amostra sem utilizar a proteômica. Todavia, os resultados gerados por ela são muitos e suas interpretações são difíceis; por conseguinte, algoritmos especializados são desenvolvidos a fim de possibilitar as análises. Algoritmos estes, que utilizam técnicas de reconhecimento de padrões probabilísticos e inteligência artificial para extrair dos resultados, informações necessárias para um determinado estudo. Devido à necessidade de avaliar e identificar cada vez mais proteínas, o estudo da proteôma (conforme explicado em 1.1) precisou ser aprofundado. Todavia, o maior desafio é justamente a complexidade do espaço de busca a ser tratado, uma vez que ele está inversamente relacionado à sensibilidade de uma ferramenta de busca. Ou seja: quanto maior o tamanho de um banco de dados (espaço de busca a ser percorrido), menor é o número de identificações obtido pela ferramenta.

O amplo aumento na diversidade de uma proteína está relacionado às suas modificações pós-traducionais, seja ela fosforilação, glicosolação, hidroxilação etc., e também à idade e à saúde das próprias células.

Com a tradução do RNA mensageiro pelo ribossomo, as proteínas podem sofrer modificações, alterando suas características estruturais e, consequentemente, a sua função. São as modificações, por exemplo, que determinam a atividade, a localização e até as interações que as proteínas terão com outras. Estas modificações podem determinar sua localização celular e fazer com que ocorra a ativação ou inativação de uma determinada função biológica.

Existem centenas de modificações pós-traducionais conhecidas e descritas em banco de dados, como, por exemplo, o *Unimod* [2].

Exemplos de modificações pós-traducionais são:

- Fosforilação: adição de um grupo de Fosfato (PO_4);
- Metilação: substituição de um átomo de hidrogênio por um grupo metil (CH_3);
- Sulfatação: adição do grupo de Sulfato;
- Formação de pontes dissulfeto: ligação entre átomos de Enxofre (S);
- Acetilação: adição de um grupo acetila (CH_3 CO); entre outras.

1.2 Enzimas proteolíticas

As enzimas são compostos orgânicos, que tem atividade intra ou extracelular e funções catalisadoras, acelerando reações químicas, que sem a presença enzimática,

teriam poucas chances de serem realizadas. Cada enzima possui uma determinada especificidade, ou seja, cada uma atua somente em alguns tipos de substratos ou sítios. A partir de características hidrofóbicas, ou hidrofílicas, ou mesmo a carga, é que tais especificidades são determinadas.

Um exemplo de enzima, cuja função é a clivagem de proteínas, é a tripsina. Esta, bastante usada neste trabalho, cliva especificamente nas ligações peptídicas, após os aminoácidos arginina e lisina (no sentido C-terminal), desde que não acompanhados de uma Prolina [32]. Ela é produzida pelo pâncreas em uma forma não ativa denominada tripsinogênio, tornando-se ativa quando alcança o duodeno[2].

A identificação de modificações pós-traducionais (PTM's) é de grande importância para o entendimento das funções proteicas. Entretanto, há muitas dificuldades para identificar as PTM's em estudos em larga, devido à baixa eficiência dos métodos apropriados. Muitas das identificações foram realizadas com a ajuda de técnicas de enriquecimentos pois, a análise das modificações requer o isolamento da proteína processada em larga escala. Um método utilizado nos dias atuais é a espectrometria de massa pois, a partir de espectrômetros de alta resolução, é possível identificar em qual resíduo encontra-se a alteração, através da sensibilidade do equipamento, e também graças a avanços recentes, como a Dissociação por Transferência de Elétron – ETD (Coon, J.J.,2009).

1.3 O fluxo da análise de dados

Para melhor conhecimento de um sistema biológico é necessário aprimorar a identificação dos padrões diferenciais que cada sistema possui. Entretanto, a proteômica *shotgun* gera uma quantidade exorbitante de informações que dificulta a análise dos dados até para os melhores especialistas, quando aplicadas às amostras complexas, como lisado de levedura[3], ou mesmo soros e tecidos.

Antes da análise dos dados gerados pelos espectrômetros de massa, um protocolo foi executado para que um maior número de dados pudesse ser identificado. Conforme mostra a Figura 1, o início está no preparo da amostra. Duas técnicas utilizadas e eficientes são: a separação por *off gel* e a tecnologia *MudPIT*, *Multidimensional Protein Identification Technology*.

[2] Tubo que liga o estômago ao intestino delgado
[3] Lisados são elementos biológicos pertencentes no citoplasma do tecido celular. Os lisados de levedura constituem seres unicelulares oriundos do Reino *Fungi*.

Após a separação, é necessário analisar as proteínas por espectrometria de massa. Essa parte é desafiadora, pois, a partir dos dados gerados pelos espectrômetros de massa, programas de computadores são utilizados para convergir a uma lista de identificações e quantificações confiáveis.

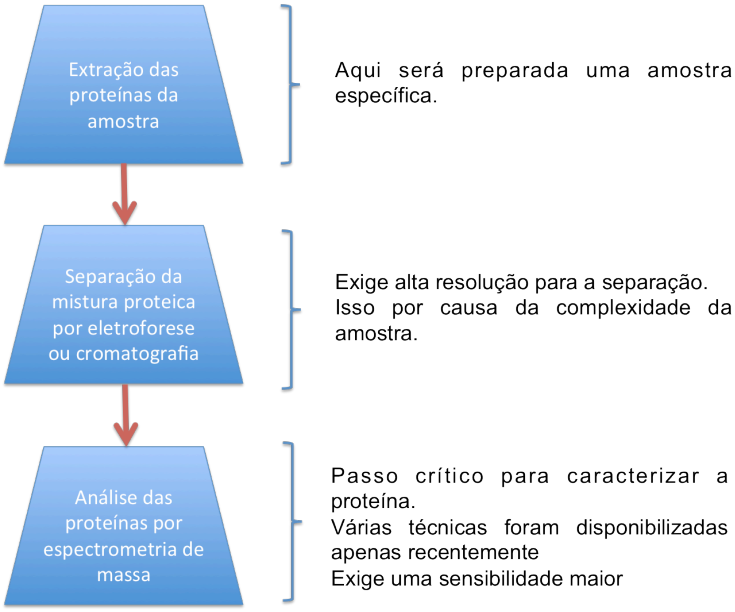


Figura 1 ***[Figura modificada a partir de [10] – Figura 3]***

A separação por *off gel* utiliza uma focalização isoelétrica que permite a separação de vários peptídeos em uma solução onde há um gel com o gradiente de pH imobilizado (estável). Por outro lado, a Cromatografia Líquida Bidimensional (LC 2D), compreendida pelo *MudPIT*, "utiliza coluna de troca iônica seguida de coluna de fase reversa diretamente acoplada ao espectrômetro de massa. A cromatografia bi-dimensional realiza-se aplicando na eluição a função degrau de aumento de concentração salina, liberando pacotes de peptídeos da coluna de troca iônica para a coluna da fase reversa (RP). Cada eluato obtido da coluna de troca iônica é posteriormente submetido ao gradiente hidrofóbico na coluna RP e, os peptídeos identificados por MS/MS." (trecho retirado de [11]).

a)  b) 

Figura 2: a – Separação por *off-gel* ; b – Tecnologia MudPIT, *Multidimensional Protein Identification Technology* (LC 2D)*[Figura retirada de http://pcarvalho.com/patternlab/mudpitsim.shtml]*

Por conseguinte, a partir do surgimento de técnicas de ionização suave (e.g., MALDI – *Matrix-assisted laser desorption/ionization*(Oishi, Y. et al.,2006) e ESI - *Electrospray Ionization*(Whitehouse, C.M. et al.,1985)), equipamentos de alto desempenho, como espectrômetro de massa de alta resolução (*e.g. Orbitrap (Makarov, A.,2000)*) e também de metodologias para fragmentação de polipeptídeos (como o ETD(Coon, J.J. et al.,2005), ou HCD), surgiu também a necessidade do desenvolvimento de novos algoritmos lógicos, a fim de analisar com a mais alta performance os dados gerados por esses equipamentos.

1.4 O uso do reconhecimento de padrões na proteômica

O Reconhecimento de Padrões, ou *Pattern Recognition*, é uma subárea da Inteligência Artificial que trata da classificação e descrição de objetos. Um sistema que envolve esta ciência está compreendido em:

- Extrair características dos objetos a serem classificados ou descritos;
- Selecionar as características mais discriminativas; e
- Construir um classificador.

De acordo com o tipo de objetos que queremos classificar, um projeto que envolve Reconhecimento de Padrões utiliza algumas abordagens em seu desenvolvimento, tais como:

- Abordagem estatística: é a mais antiga de todas, conhecida como "Teoria da Decisão". Ela assume que as classes são definidas a partir de modelos probabilísticos;

- Abordagem neural: procura determinar um mapeamento ótimo inspirado na rede neural do cérebro, contendo neurônios que se interligam, representando as ligações sinápticas; e
- Abordagem difusa: esta última é uma abordagem que leva em conta o grau de incerteza das características, que muitas vezes ficam ocultas. Ela utiliza a Teoria dos Conjuntos Difusos, a qual define o grau de pertinência que cada elemento contém.

Na proteômica, o reconhecimento de padrões é usado para possibilitar a análise e interpretação dos resultados. Contudo, há outros fatores limitantes que faz com que tal técnica seja mais desenvolvida, como o custo de equipamentos e reagentes utilizados na amostra, o número de parâmetros utilizados pela mesma, o enorme número de variabilidade entre uma amostra e outra, a própria limitação da reprodução das metodologias proteômicas para detecção e quantificação de proteínas em larga escala e, principalmente, o não reconhecimento de funções densidade de probabilidade (pdf) capazes de representar a distribuição de probabilidade do nível de expressão de determinada proteína nos estudos em questão.

Com o intuito de minimizar tais fatores, algoritmos de IA são cada vez mais refinados, adaptando-se às amostras a serem analisadas. Com isso, a metodologia de *Machine Learning*, tem como objetivo aprimorar o número de identificações de proteínas de cada espectro em questão, fazendo com que a sensibilidade da busca fique melhor.

Ao final de um determinado experimento, uma lista de espectros precisará ser analisada por algoritmos responsáveis por assimilar quais proteínas são as mais plausíveis na identificação de uma determinada sequência. É neste momento, que as metodologias de inteligência artificial aparecem, distinguindo para um determinado caso, qual é a mais apropriada.

1.4.1 Algoritmos para identificação de peptídeos analisados por espectrometria de massa em tandem

Na proteômica, existem três metodologias canônicas capazes de sequenciar peptídeos analisados por espectrometria de massa em tandem. São elas: *de novo sequencing*, *sequence tag search* e o *peptide spectrum match* – PSM. Cada um destes métodos apresentam vantagens e desvantagens.

O *de novo sequencing* está exemplificado na Figura 3. A principal vantagem é não requerer um banco de dados de proteínas para realizar identificações. Ele é usado, em sua grande maioria, quando o estudo envolve organismo(s) não sequenciado(s). Em linhas gerais, o algoritmo desenvolve um grafo, cujos *nós* são as massas relacionadas a cada aminoácido encontrado. Um caminho ótimo é traçado a fim de obter a sequência peptídica mais apropriada. A desvantagem desta técnica é que está mais propícia a erros, justamente porque os grafos possuem vários caminhos ótimos, ficando difícil, muitas vezes impossível, definir qual o caminho correto. O *PepNovo* [1], *pNovo+*[16] e o *Peaks*[33], são exemplos de ferramentas que utilizam esta metodologia.

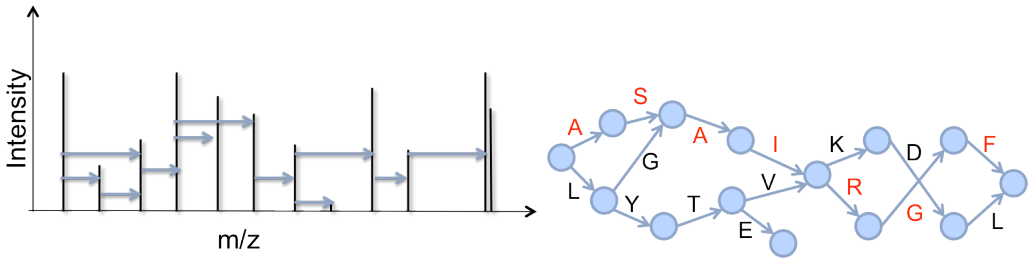


Figura 3: Metodologia de sequenciamento de espectros: *de novo sequencing*

O *sequence tag search*, demonstrado na Figura 4, define uma classe de algoritmos que obtém *sequence tags* de aproximadamente dois a quatro aminoácidos, e utilizam os mesmos em um esquema de votação através de um banco de dados para selecionar a sequência que melhor explica os *tags*. A vantagem é de ser tolerante a modificações pós-traducionais e a mutações não especificadas *a priori*. Entretanto, se a proteína estiver presente no banco de dados, a sensibilidade desta metodologia não é tão eficaz quanto à do *peptide sequence matching*, descrita a seguir. O *GuttenTag* (Tabb, D.L., Saraf, A., and Yates, J.R., III,2003) é uma ferramenta de busca que utiliza esta técnica implementada em seu núcleo.

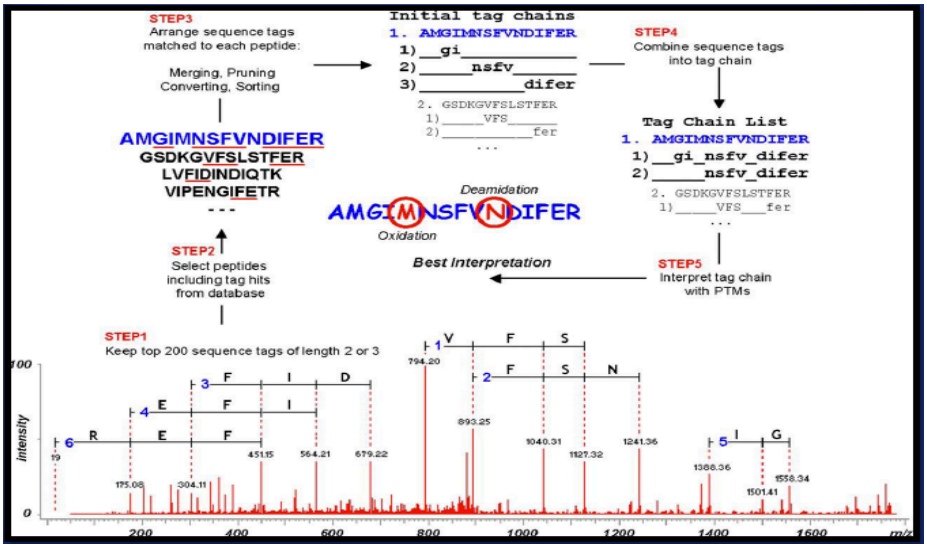


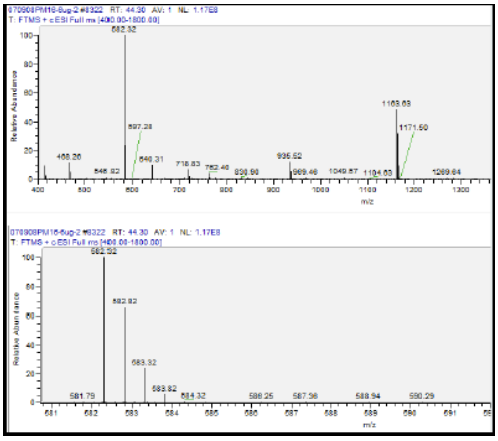
Figura 4: Representação da *Sequence Tag Search* (Na S. et al., MCP, 2008)

E por último, os algoritmos do tipo *peptide spectrum match*, ou PSM, que são considerados como padrão ouro. O PSM é a técnica mais sensível, desde que a sequência esteja depositada no banco de dados. Este, por sua vez, é a principal desvantagem desta técnica, pois há a necessidade de ter um conjunto de sequências armazenadas para que se possa fazer a identificação. E, além disso, devem-se especificar *a priori* todas as modificações pós-traducionais a serem consideradas.

Existem vários bancos de dados de sequências proteicas publicamente disponíveis. Os principais são:

- *SWISS-PROT*: é um banco de dados de anotações de sequências proteicas. Contém informações adicionais na função proteica, assim como conhecidas modificações pós-traducionais;
- *TrEMBL*: contém a maioria das traduções das entradas das sequências nucleotídicas que ainda não foram integradas ao SWISS-PROT;
- *PIR-International*: banco de dados de anotações de sequências proteicas;
- *NCBInr*: contém sequências de DNA do GenBank, SWIS-PROT e do PIR;
- *UniProt*: essa é uma nova proposta de banco de dados. Ele junta os bancos SWISS-PROT, TrEMBL e PIR.

A Figura 5 esquematiza o funcionamento da metodologia PSM.

a) 

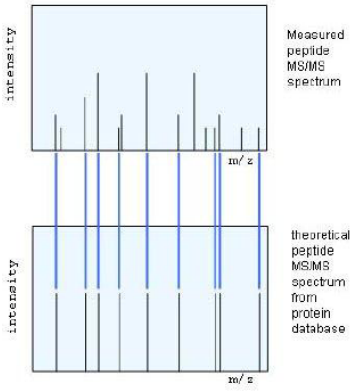
b) 

Figura 5: Metodologia *peptide spectrum match* - PSM. a-Espectros experimentais; b-*Matching* do espectro teórico (na parte inferior) com o espectro experimental (na parte superior). ***[Figura retirada de "http://proteomesoftware.com"]***

Pode-se citar como exemplos de ferramentas que utilizam PSM:

- *SEQUEST*: Este *software* avalia as sequências proteicas de um banco de dados para computar a lista de possíveis peptídeos candidatos. Esta avaliação é inicialmente realizada selecionando os peptídeos que tem a massa mais próxima do espectro analisado. Para cada candidato, o *SEQUEST* projeta um espectro de massa de íon dissociado (ou MS2) e compara com o espectro de massa experimental através da técnica de correlação cruzada[4]. O candidato com o melhor *matching* é reportado como a melhor identificação para aquele determinado espectro. [37]
- *Mascot*: é uma ferramenta proprietária da *Matrix Science*. Assim como o *SEQUEST*, ele possui um algoritmo proprietário e é, portanto, desconhecido para realizar a comparação de espectros teóricos com experimentais.[27][34]
- *Andromeda*: Esta ferramenta está integrada no ambiente computacional *MaxQuant* e utiliza a lógica PSM como principal motor em sua busca, analisando padrões complexos de modificações pós-traducionais. [17]

1.5 Espectrômetro de massa

A LC 2D elui peptídeos para o espectrômetro de massa com um gradiente de solvente orgânico, e este por sua vez, gera frequentemente espectros de massa em tandem. Ao final do experimento, haverá uma lista de espectros a serem analisados.

[4] *Cross-correlation*, ou correlação cruzada, é uma medida de similaridade de dois sinais em função de um atraso aplicado a um desses sinais.

O espectrômetro de massa é um aparelho que ioniza moléculas para a forma gasosa e separa as mesmas de acordo com sua relação massa/carga (*m/z*). A Figura 6 esquematiza o funcionamento do espectrômetro, que procede da seguinte forma: primeiramente o analito é ionizado em uma fonte que poderá ou não, estar em um ambiente de baixa pressão (abaixo de 1 atm.). Chegando ao analisador, este irá separar os íons de acordo com a razão *m/z* deles, para aí sim, obter os perfis de espectros de massa, ou MS1. Um outro processo é a obtenção dos espectros de massa em tandem, ou MS2, os quais, íons serão previamente selecionados de acordo com sua razão *m/z*, e submetidos à fragmentação. Por fim, o detector, juntamente com uma ferramenta de proteômica computacional, compõem a última etapa deste processo, analisando a corrente iônica originária da neutralização do íon do analito, através da intensidade do sinal gerado no espectro de massa.

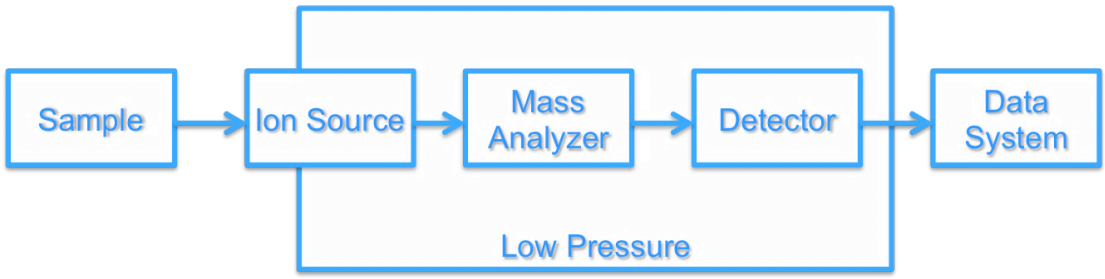
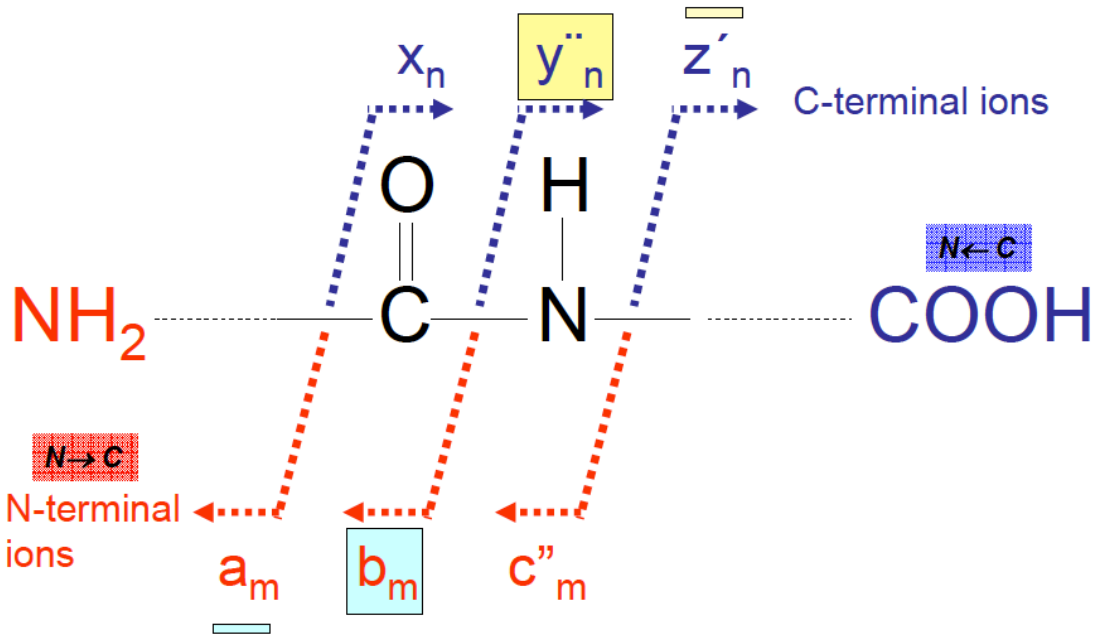


Figura 6: Fluxograma das etapas de um espectrômetro de massa ***[Figura modificada a partir de "Chemistry in action! 51 – Figura 2a"]***

Conforme apresentado anteriormente, é no Analisador que o espectro começa a ser formado, uma vez que o espectrômetro transmite energia para os peptídeos causando a fragmentação entre os aminoácidos. Este é processo é denominado *cell collision*. A partir de então, os perfis de espectros são elaborados contendo a razão m/z dos íons analisados. Estes, por sua vez, possuem tipos para serem interpretados, que são: *a*, *b*, *c*, *x*, *y* e *z*, conforme visualizado na Figura 7.



Roepstorff, P., and Fohlmann J. *Biomed. Mass Spectrom.* **11**, 601 (1984)

Jhonson R. S., Martin S. A., Biemann K., *Int. J. Mass Spectrom. Ion Processes.* **86**, 137 (1988).

Figura 7: Representação esquemática dos tipos de fragmentação entre os aminoácidos que ocorrem pelo processo de colisão celular.

Os íons mais comuns nos espectros de peptídeos fragmentados por CID são os do tipo *b* e *y*. Os do tipo *b* são interpretados ao ler-se o espectro da esquerda para a direita, e a distância entre cada pico é dada de acordo com a massa molecular de um determinado aminoácido, como demonstrado na Figura 8.

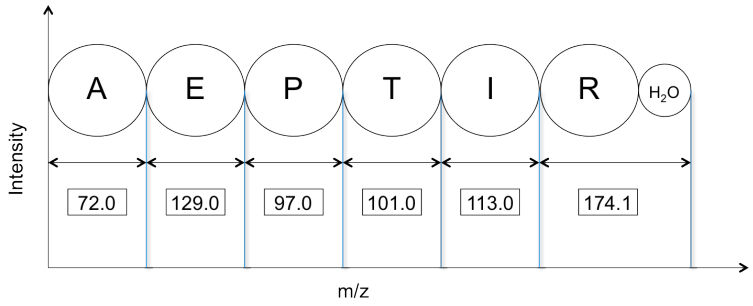


Figura 8: Representação esquemática da fragmentação tipo *b* formando a sequência AEPTIR

Já os íons do tipo *y* são obtidos fazendo a leitura do espectro na ordem inversa, ou seja, da direita para a esquerda, como demonstrado na Figura 9.

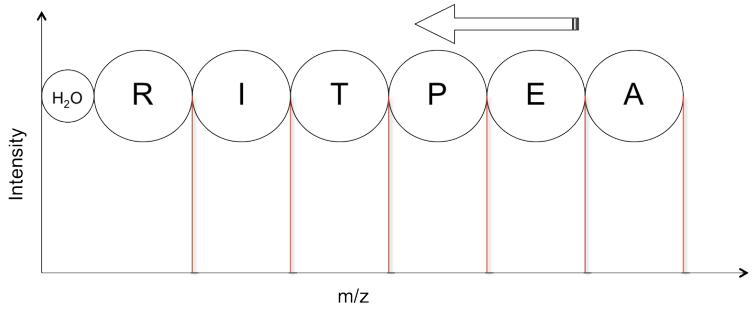


Figura 9: Representação esquemática da fragmentação tipo *y* formando a sequência AEPTIR

Porém, os íons apresentam intensidades diferentes de um pico ao outro, e todos eles estão contidos em um mesmo espectro, o qual possui também, ruídos consideráveis. Logo, a dificuldade do interpretador é analisar o que é válido e quais aminoácidos estão presentes naquele espectro, justamente para compor a sequência peptídica. A Figura 10 demonstra um espectro teórico com os íons tipo *b* e *y* e um espectro experimental, onde existem, além dos picos dos íons tipo *b* e *y*, ruídos para dificultar a interpretação.

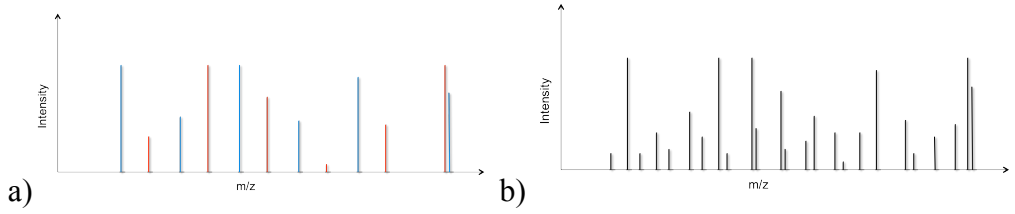


Figura 10: a-Espectro de massa teórico com íons tipo *b (na cor azul)* e *y (na cor vermelha)*. b- Espectro de massa experimental com ruídos, além dos íons tipo *b* e *y*.

1.5.1 A ferramenta de busca

Para que as identificações possam ser realizadas, é necessário fazer a busca utilizando uma ferramenta de proteômica computacional. Como explicado em [1.4.1], existem três metodologias canônicas capazes de fazer o sequenciamento dos espectros, porém a técnica proposta nessa dissertação será aquela tida como padrão ouro, a PSM. Uma ferramenta de busca PSM funciona conforme descrito a seguir: primeiramente, é alimentada por uma coleção de espectros de massa. A partir daí, haverá a comparação entre os espectros teóricos, obtidos através do banco de dados de peptídeos, e os espectros experimentais, gerados pelo espectrômetro de massa de peptídeos que apresentem massa dentro de uma tolerância previamente especificada. Na *Search Engine* será verificado qual peptídeo mais se assemelha ao espectro experimental, de acordo com uma determinada métrica da ferramenta de busca.

Aquele que tiver maior similaridade é o que será o *peptide spectrum match*. A Figura 11 demonstra o fluxo de dados de uma ferramenta de busca que utiliza a metodologia PSM.

É na *Search Engine* que está implementado técnicas de Inteligência Artificial – IA, como Reconhecimentos de Padrões [1.4], uma vez que quanto mais aprimorada está esse núcleo da ferramenta computacional, mais sensível será a identificação de peptídeos para uma determinada organela. Mais adiante, será explicada a técnica de IA implementada no *software* desenvolvido durante a confecção desta dissertação.

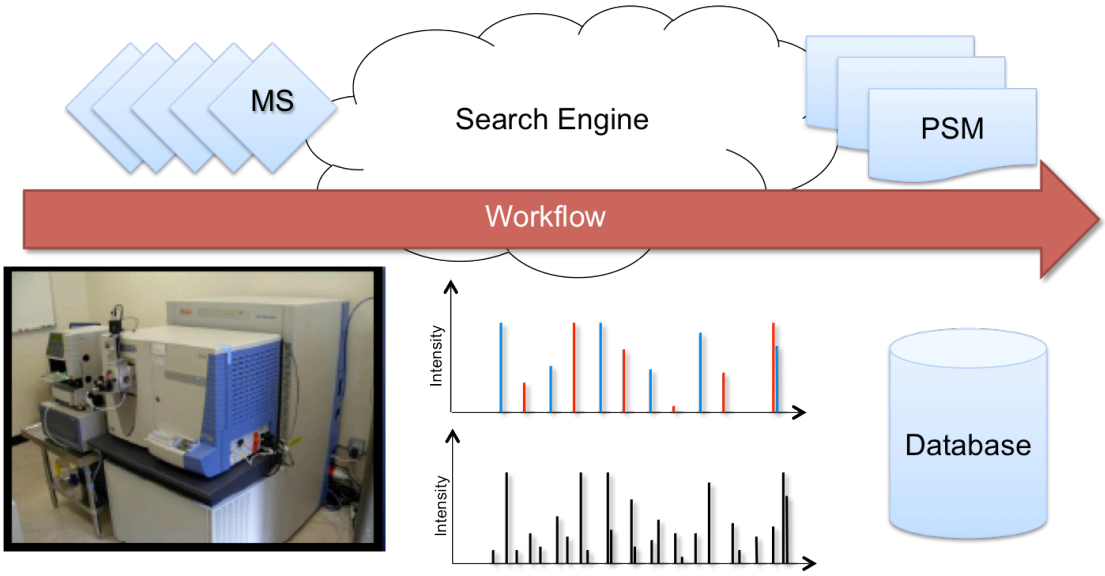


Figura 11: Fluxo de dados de uma ferramenta de busca utilizando *peptide spectrum match* – PSM. A ferramenta é alimentada por uma coleção de espectros de massa (MS) e fará a comparação dos espectros teóricos (obtidos a partir do banco de dados) com os espectros experimentais. O espectro teórico que tiver mais semelhança com o experimental é o que vai ser o espectro PSM.

Contudo, para que se possa realizar um estudo com dados proteômicos, é necessário fazer um pós-processamento dos dados. Isso ocorre porque o estudo necessita que os resultados convirjam para uma lista de identificações confiáveis, excluindo, assim, aquelas cujo *score*[5] é pobre.

Existem várias ferramentas que executam essa filtragem, como o *DTASelect* que organiza e filtra as identificações do *SEQUEST*, reduzindo o tempo necessário para interpretar os resultados de cada amostra [18]; e também o *Percolator*, que utiliza uma máquina de aprendizado semi-supervisionado, a fim de melhorar a discriminação entre as corretas e as incorretas identificações de espectro [31].

[5] O *score* é calculado a fim de obter o grau de afinidade que um espectro teórico tem em relação ao experimental. O cálculo é realizado de acordo com métricas definidas em cada ferramenta de busca.

Todavia, nesta dissertação foi utilizada o *Search Engine Processor* (SEPro)[12]. Um *software* desenvolvido com o objetivo de fazer esse filtro estatisticamente, detalhando os peptídeos que foram encontrados e também os espectros teóricos que tiveram maior similaridade com os experimentais. Ele utiliza um classificador Bayesiano que utiliza espectros, peptídeos e uma lógica proteica que trata o resultado e converge os dados para uma lista de identificações confiáveis.

1.6 Complexidade dos espaços de busca

Conforme explicitado em [1.1], quanto maior o tamanho do espaço de busca, ou seja, quanto maior o tamanho do banco de dados de proteínas, menor é a sensibilidade de uma ferramenta de busca.

Um estudo de uma microbiota de bactérias, por exemplo, onde não se sabe quais os tipos de bactérias estão ali presentes, levará naturalmente, na concatenação de vários bancos de sequências relacionados a esses microrganismos, a fim de descobrir àquelas pertencentes na microbiota. Um outro exemplo é o estudo de veneno de serpentes. Nele, é muito comum estar em busca de peptídeos naturais, logo não se pode impor, em uma ferramenta de busca, uma condição tríptica, uma vez que o objetivo é encontrar sequências em sua forma natural. A consequência disto está na quantidade de peptídeos que irá ser encontrada. Estes foram apenas dois exemplos de muitos outros existentes, os quais o tamanho do banco de dados de proteínas irá ser enorme e, consequentemente, a sensibilidade de uma ferramenta de busca ao realizar as identificações em um espaço desses irá ser baixa.

Figura 12: A microbiota de bactérias e os venenos de cobra são exemplos de que o espaço de busca tem uma complexidade enorme.
[figuras retiradas de http://www.probisearch.com/?page_id=2721&lang=en http://correio.rac.com.br/_conteudo/2012/11/capa/nacional/7585-veneno-de-cobra-e-testado-contra-o-cancer.html]

1.6.1 Tipos de espaços de busca

Ao realizar uma busca de proteínas, a ferramenta computacional irá procurar nos espaços os espectros que terão maior similaridade com aquele em questão. Quanto mais sequências de peptídeos candidatos a serem procurados, maior é a chance de errar. O tamanho do espaço de busca é uma função do número de sequências proteicas, da enzima utilizada e número de PTM's consideradas [1.2]. A tripsina foi a enzima utilizada neste trabalho para realizar a clivagem, e ela, por sua vez, realiza a quebra sempre após os aminoácidos arginina e lisina, representados respectivamente por *R* e *K*.

No espaço tríptico, a tripsina fará com que a sequência complexa de peptídeos seja clivada em sequências menores, havendo um conjunto de aminoácidos compreendidos entre *R* ou *K*. Já no espaço semi-tríptico, as sequências menores terão sempre a especificidade em um terminal tríptico. Isso fará com que o espaço aumente razoavelmente. E finalmente, no que definimos de espaço não-tríptico, a clivagem deverá ser realizada em todas as possíveis ligações peptídicas (e combinações) de cada sequência complexa. Dessa forma, o tamanho efetivo do espaço de busca será consideravelmente maior, tornando-se este espaço o mais complexo a ser tratado.

Baseado no exemplo de peptídeos naturais mencionado anteriormente, a seguir será explicado como um espaço não tríptico, cuja enzima proteolítica não clivou as sequências peptídicas exatamente nos aminoácidos esperados, torna-se complexo rapidamente.

Para melhor exemplificar, considere um banco de dados contendo apenas uma sequência peptídica: ***ARSPTEGLKID***, e que a ferramenta de busca tenha uma restrição em que só aceite como resultados peptídeos compreendidos em mais de quatro aminoácidos. No espaço de busca tríptico o tamanho efetivo é de apenas um, conforme demonstrado na Tabela 1. Quando se aumenta um pouco o espaço, chegando ao semi-tríptico, o tamanho já tem um salto grande, indo para onze peptídeos. Todavia, o espaço mais crítico é o não-tríptico, cujo tamanho foi de um para 34 peptídeos.

Tabela 1: Tamanho efetivo de cada espaço de busca gerado a partir da clivagem da enzima proteolítica tripsina.

Espaço Tríptico	Espaço Semi-tríptico	Espaço Não-tríptico
AR.*SPTEGLK*.ID	SPTEGLKID, SPTEGLKI, SPTEGLK, SPTEGL, SPTEG, SPTE, ~~SPT~~, ~~SP~~, ~~S~~, ~~K~~, ~~LK~~, ~~GLK~~, EGLK, TEGLK, PTEGLK, RSPTEGLK, ARSPTEGLK,	SPTEGLKID, SPTEGLKI, SPTEGLK, SPTEGL, SPTEG, SPTE, ~~SPT~~, ~~SP~~, ~~S~~, PTEGLKID, PTEGLKI, PTEGL, PTEG, ~~PTE~~, ~~PT~~, ~~P~~, TEGLKID, TEGLKI, TEGLK, TEGL, ~~TEG~~, ~~TE~~, ~~T~~, EGLKID, EGLKI, EGLK, ~~EGL~~, ~~EG~~, ~~E~~, GLKID, GLKI, ~~GLK~~, ~~GL~~, ~~G~~, LKID, ~~LKI~~, ~~LK~~, ~~L~~, ~~KID~~, ~~KI~~, ~~K~~, ~~ID~~, ~~I~~, ~~D~~, ~~K~~, ~~GLK~~, EGLK, TEGLK, PTEGLK, RSPTEGLK, RSPTEGL, RSPTEG, RSPTE, RSPT, ~~RSP~~, ~~RS~~, ~~R~~, ARSPTEGLK, ARSPTEGL, ARSPTEG, ARSPTE, ARSPT, ARSP, ~~ARS~~, ~~AR~~, ~~A~~
1 peptídeo	**11 peptídeos**	**34 peptídeos**

Os peptídeos grifados na Tabela 1 não satisfazem ao exemplo de condição da ferramenta de busca de apenas aceitar peptídeos que tenham mais de quatro aminoácidos em sua sequência. E como se pode observar, do espaço tríptico para o não-tríptico houve um aumento de mais de 3000% fazendo com que este último tornasse complicado de ser tratado. Nesta dissertação está sendo considerado espaço de busca complexo aquele cujo tamanho efetivo é superior a 50 milhões de peptídeos.

Considerando agora um exemplo real, a Figura 13 representa os tamanhos proporcionalmente distribuídos do banco de dados da *Escherichia Coli*[6] com aproximadamente quatro mil proteínas. Como se pode notar, o espaço tríptico é o menor, representado apenas por um círculo do tamanho de um ponto. Por outro lado, o maior espaço efetivo compreende um tamanho com mais de 63 milhões de peptídeos.

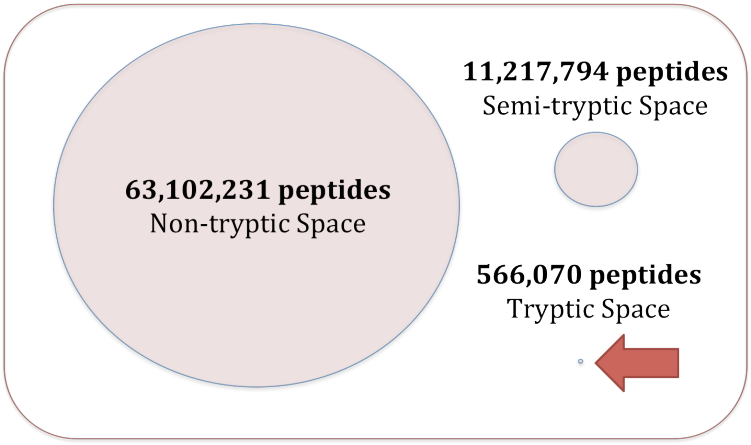


Figura 13: Tamanho efetivo gerado a partir de uma banco de dados da *E. coli* de aproximadamente quatro mil proteínas

Realizando uma busca nesses espaços apresentados anteriormente, utilizando o método de fracionamento HCD, em um espectrômetro de massa Orbitrap *Velus* e uma cromatografia de aproximadamente duas horas, obteve-se o resultado apresentado na Figura 14. Como se pode perceber, no espaço semi-tríptico, obteve-se o maior número de identificações de espectros. Isso ocorre porque vários peptídeos são clivados durante o processo de ionização, fazendo com que aquele que originalmente era tríptico, agora, se torne semi-tríptico. Mas é no espaço não-tríptico que a sensibilidade da ferramenta de busca decai bruscamente. Foram encontrados 1.352 espectros – cerca de 50% a menos que no espaço semi-tríptico – em um espaço de tamanho efetivo superior a 63 milhões de peptídeos. Isso ocorre, pois, se o espaço de busca aumenta, o número de candidatos para cada espectro também aumenta. Logo crescem as chances de erro, pois a taxa de falsos positivos é baixa, de apenas 1%.

[6] *Escherichia Coli*, ou *E.coli*, é uma bactéria que juntamente com o *Staphylococcus aureos* é a mais comum no ser humano. As primeiras evidências foram relatadas pelo alemão Theodor Escherich, em 1885. [29]

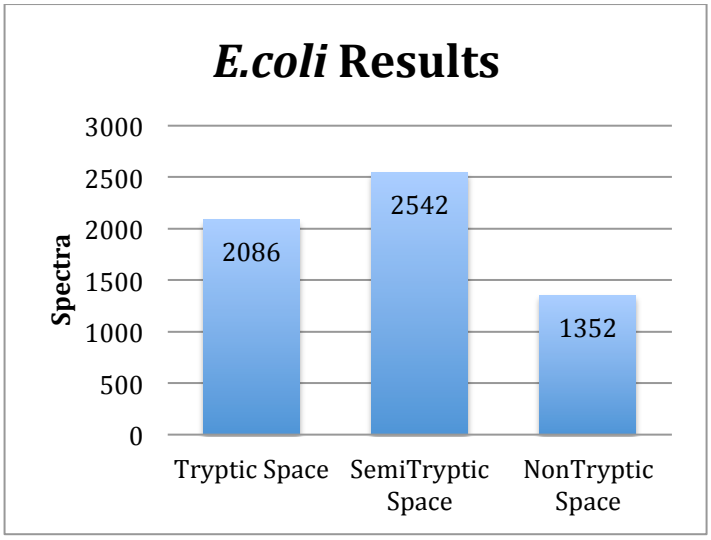


Figura 14: Resultado obtido na busca realizada nos espaços tríptico, semi-tríptico e não-tríptico utilizando o banco de dados *E.coli* com aproximadamente quatro mil proteínas.

2 Objetivos

- Criar uma ferramenta capaz de identificar peptídeos pela comparação de seu espectro de massa em tandem com teóricos gerados a partir de um banco de dados de sequências.
- Gerar uma lógica para sequenciar o primeiro aminoácido de espectros de massa de peptídeos marcados com fenil-isotiocianato (PITC).
- Acoplar a lógica PITC na ferramenta de busca para redução do espaço de busca.

3 Justificativa

Os *softwares* de bioinformática que tem como objetivo identificar padrões são computacionalmente custosos; além da demora, a chance de obter um resultado errôneo aumenta com o tamanho do espaço, fazendo com que algoritmos de filtragem estatística (e.g., DTASelect(Cociorva, D., Tabb, L., and Yates, J.R.,2006), SEPro (Carvalho, P.C. et al.,2012) etc.) sejam mais estringentes, a fim de atingir resultados com um *False Discovery Rate* (FDR) estabelecido. Através da marcação química, denominada PITC, ou fenil-isotiocianato, o íon do tipo *b1* pertencente à cadeia peptídica formadora da proteína, torna-se mais intenso. Uma lógica capaz de sequenciar o primeiro aminoácido considerando a intensidade deste íon foi desenvolvida, fazendo com que a complexidade do espaço de busca diminua, reduzindo também, tanto o tempo de execução de busca, quanto, principalmente o número de falsos positivos.

Este trabalho descreve a primeira ferramenta de busca capaz de considerar a característica do PITC em seus algoritmos de busca. Denominada *Spectrum Identification Machine* – SIM, ela tem em seu núcleo, tanto a busca considerando a marcação PITC, quanto sem a marcação, conforme as demais ferramentas (e.g., SEQUEST). Isso é fundamental, pois apesar do algoritmo ser robusto o suficiente para identificar espectros de peptídeos marcados, o mesmo também identifica peptídeos não marcados com fenil-isotiocianato, obtendo melhores resultados que outros programas como Andromeda, módulo de busca do MaxQuant, que também realiza buscas através da espectrometria de massa.

4 Metodologia

A partir dos objetivos apresentados na seção 2, o qual pretende-se aumentar a eficiência da técnica tida como padrão ouro na proteômica, a *peptide spectrum match - PSM*, em espaços de busca complexos, uma nova estratégia computacional e experimental é demonstrada nessa dissertação.

O primeiro passo foi desenvolver uma nova ferramenta de busca utilizando PSM. Esta foi desenvolvida, uma vez que os códigos fonte dos *softwares* tradicionais no ambiente proteômico não estão disponíveis. Logo, eles não poderiam ser utilizados para fazer experimentos mais aprofundados, pois precisaria de um controle preciso da ferramenta para obter uma performance maior. E, aproveitando que estávamos desenvolvendo uma nova ferramenta, denominada *Spectrum Identification Machine*, embutimos também novos conceitos desenvolvidos por pesquisadores brasileiros. Por exemplo, uma ferramenta tradicional como o SEQUEST, nos espectros teóricos, os íons tipo *b1* e *y1* apresentam intensidades constantes. Por outro lado, ao dar pesos a regiões diferentes do espectro, a eficiência da busca será maior, como demonstrou em seu artigo publicado na *Journal of Proteome Research* – JPR o Junqueira, M. *et al*.[24]. De acordo com os espectros da Figura 15, o Professor Junqueira apresentou a importância dos pesos a regiões diferentes, dando chances similares a picos com baixa intensidade. O SIM implementa esta metodologia e consegue convergir a pesos ótimos de acordo com uma técnica de *Machine Learning*, ou seja, a eficiência dos espectros teóricos apresentada pela ferramenta de busca criada é mais eficiente do que a metodologia padrão, utilizada por muitas ferramentas de busca.

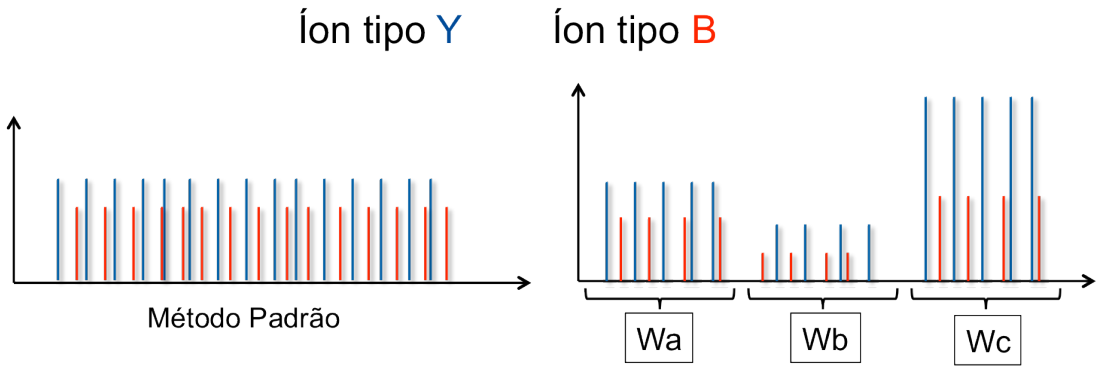


Figura 15: O aumento da eficiência da busca dar-se pela distribuição de pesos a regiões diferentes ao longo do espectro. Enquanto no gráfico da esquerda temos o espectro de massa teórico representando o método padrão, o gráfico da direita temos os pesos (*Wa*, *Wb* e *Wc*) atribuídos a certas regiões do espectro.

4.1 Spectrum Identification Machine

Explorar o campo proteômico por meio de técnicas computacionais foi o desafio atribuído nessa dissertação, e, nesse âmbito, foi desenvolvido o *Spectrum Identification Machine* – SIM, uma ferramenta que tem como propósito aumentar a sensibilidade na identificação de peptídeos a partir da comparação de espectros teóricos e espectros experimentais. Porém, com as modificações pós-traducionais da cadeia proteica, o espaço de busca aumenta exponencialmente. Através da marcação fenil-isotiocianato (PITC) [4.2], pôde-se reduzir o espaço de busca.

Nesta seção, será detalhado o funcionamento do SIM, como ele tirou proveito da marcação PITC e o detalhamento dos métodos por ele utilizados.

4.1.1 Parâmetros

Antes de começar a detalhar o SIM, é necessário explicar os parâmetros por ele utilizados para que sua performance possa atingir o máximo possível.

Para que as amostras pudessem ser analisadas, foi definido um intervalo de tolerância, o qual é medido em *ppm (parte por milhão)*, utilizado tanto no MS1 quanto no MS2. Dessa forma pode-se obter amostras dentro de um espaço de *ppm* previamente definido, o qual é uma medição da acurácia do espectrômetro. O número mínimo e máximo de aminoácidos[7] em cada sequência peptídica também precisou ser parametrizado. Assim, só poderão ser analisadas sequências que estejam entre um

[7] Aminoácido é uma molécula orgânica que contém um grupo amina, um outro grupo carboxila e uma cadeia lateral específica para cada molécula. Os principais átomos contidos em um aminoácido são: carbono, hidrogênio, oxigênio e nitrogênio.

intervalo de tamanho pré-estabelecido. Definiu-se também as modificações que uma sequência proteica poderia sofrer. No SIM, está pré-definido a "carbamidometilação de cisteína"[8], cujo seu *DeltaMass* é 57.02146 Da; e a "oxidação de metionina"[9], cujo *DeltaMass* é de 15.9949 Da. A enzima proteolítica escolhida para clivar a sequência proteica foi a Tripsina, uma vez que ela sempre cliva no grupo carboxilo da Arginina ou da Lisina. Outros parâmetros foram definidos, mas serão explicados mais adiante.

Uma vez configurado os parâmetros, estes serão serializados[10], para que, numa próxima execução do programa, o tempo de resposta ao início do processo seja menor. O SIM agora criará um conjunto de informações, chamado dicionário, com as massas residuais dos aminoácidos, além da massa monoisotópica[11] de algumas moléculas. Também serão adicionados no dicionário o *DeltaMass* das modificações que a cadeia proteica poderá sofrer. Este dicionário facilita no instante da busca do aminoácido, uma vez que o acesso do dado tem um tempo de complexidade $O^{12}(1)$. A Tabela 2 mostra os respectivos dados utilizados no *software*:

Tabela 2: Aminoácidos com suas respectivas massas em Da.

Aminoácido / Molécula	Descrição	Massa
G	Glicina	57,0214637
A	Alanina	71,0371138
S	Serina	87,0320284
P	Prolina	97,0527638
V	Valina	99,0684139
T	Treonina	101,047678
C	Cisteína	103,009185
I	Isoleucina	113,084064

[8] Carbamidometilação de cisteína é uma modificação que previne que as pontes dissulfeto quebradas na síntese proteica não voltem a ser ligadas.
[9] Oxidação de metionina é o aumento da carga elétrica do aminoácido metionina, onde as moléculas deste aminoácido perdem elétrons na reação química.
[10] Serializar um objeto é colocar os valores nele contidos juntamente com suas propriedades de certa maneira que fiquem em série, daí o nome serial. Dessa forma, um objeto serializado terá os privilégios para que ele possa ser gravado em disco ou mesmo transmitido pela rede.
[11] Massa monoisotópica corresponde à soma das massas dos átomos de uma molécula utilizando a massa do isótopo mais abundante. Para a grande maioria dos compostos orgânicos, a massa monoisotópica corresponde à massa do isótopo mais abundante.
[12] Complexidade computacional é um ramo da teoria da computação que se concentra em classificar problemas de acordo com sua dificuldade. Quando o acesso à informação é de forma imediata, sem a necessidade de resolver cálculos aprofundados, dizemos que o tempo é $O(1)$.

L	Leucina	113,084064
N	Aspargina	114,042927
D	Ácido Aspártico	115,026943
Q	Glutamina	128,058578
K	Lisina	128,094963
E	Ácido Glutâmico	129,042593
M	Metionina	131,040485
H	Histidina	137,058912
F	Fenilanina	147,068414
U	Selenocisteína	150,95364
R	Arginina	156,101111
X ou J	Leucina ou Isoleucina	113,08406
Y	Tirosina	163,063329
W	Triptofan	186,079313
O	Pirrolisina - O 22º aminoácido	255,166692
H	Hidrogênio	1,007825032
O	Oxigênio	15,99491462
C	Carbono	12
N	Nitrogênio	14,00307401
NH_3	Amina	17,0265491
CO	Monóxido de Carbono	27,99491462
H_2O	Água	18,01056469
B	Aspargina ou Ácido Aspártico	114,042927
Z	Ácido Glutâmico	128,058578

4.1.2 Linguagem de programação

A esquematização do *software* é parte fundamental para futuras manutenções e aprimoramentos na lógica da programação. Pensando nisso, foi fundamental a escolha da metodologia a ser seguida e, o paradigma da orientação a objetos torna-se o código fonte cada vez mais robusto e organizado.

O SIM está seguindo o padrão *Model-View-Controller* (MVC), que é um padrão bastante difundido na área de desenvolvimento de sistemas. Isso, porque a ferramenta de busca poderá ser executada tanto em linha de comando, quanto em uma

interface gráfica (GUI), facilitando o manuseio do *software* por usuários e também por *clusters* de processamento.

4.1.2.1 MVC – *Model-View-Controller*

Com o intuito de separar a lógica de negócio da apresentação, foi criado o padrão *Model-View-Controller*, ou simplesmente MVC. Ele foi criado a partir da necessidade de organizar o código de sistemas bastante complexos, tornando-se muito viável a separação dos dados da aplicação com a sua visualização.

Dessa forma, o MVC é compreendido por três camadas, conforme pode-se observar na Figura 16:

1. ***Model***: Responsável por reunir as informações que mostram o estado de um componente, além de informar para seus observadores sobre as mudanças ocorridas nos dados. É no *model* que se gerencia e definem-se as classes de domínio.
2. ***View***: É a parte da aplicação que interage com o usuário. É na *view* que haverá a integração do *model* e a especificação da maneira como os dados serão apresentados ao usuário.
3. ***Controller***: Responsável pelo tratamento de eventos, ou seja, é no *controller* que as informações e/ou eventos do usuário, realizados na *view*, serão capturados e processados para que o *model* seja modificado. Ele é responsável também por validar e filtrar a entrada de dados realizada pelo usuário. (trecho retirado de [28])

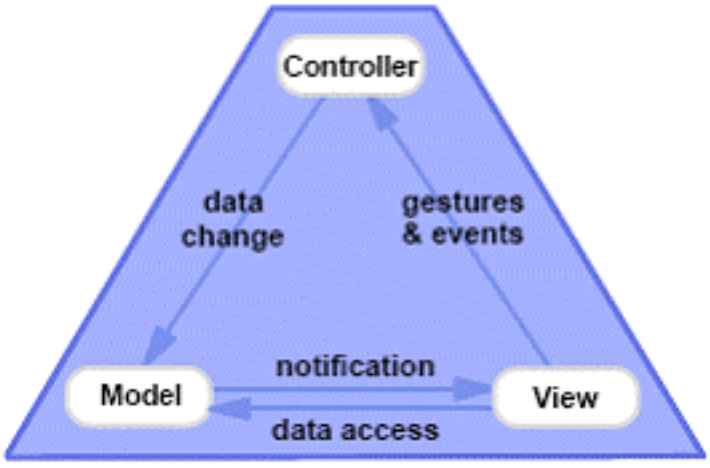


Figura 16: Interação dos componentes do MVC [*figura retirada de [28]*]

Com o padrão MVC estabelecido, o próximo passo foi escolher a linguagem de programação a ser utilizada no desenvolvimento da ferramenta de busca. Pelas

facilidades atribuídas à linguagem, assim como características funcionais pertencentes somente na plataforma .NET, e que facilita muito a otimização do código, o C# foi escolhido, uma vez que, ele é amplamente difundido, facilitando na integração de fóruns de dúvidas, fundamentais no ambiente de desenvolvimento. Características estas, como, por exemplo, o *LINQ*, que é bastante útil no manuseio de objetos, atribuindo valores de forma rápida e realizando consultas aprimoradas.

4.1.3 Banco de dados de peptídeos

Após criar o dicionário de massas residuais, o SIM preparará seu banco de dados de peptídeos. Este banco será composto a partir de sequências proteicas pré-estabelecidas. A ferramenta possui um *parser*[13] o qual lê o arquivo do banco de dados com as sequências e cria um dicionário com todos os peptídeos encontrados. Para encontrar tais peptídeos uma digestão é realizada. Digestão esta que fará a clivagem das sequências proteicas de acordo com a enzima proteolítica escolhida neste trabalho; a tripsina. Ela fará a clivagem após os aminoácidos *K* e *R*. Porém, a digestão pode ser incompleta, dando origem a *missed cleavages*. Podem existir inúmeros *missed cleavages* dentro de uma mesma sequência proteica. Para isso, o SIM determina um parâmetro onde o número máximo deste efeito é configurado. Sabendo-se disto, a digestão da proteína é realizada de acordo com a especificidade da enzima. A especificidade enzimática é a capacidade que cada enzima tem de agir sobre um determinado substrato. Um substrato é um composto químico o qual sofre a reação catalisada da enzima.

No SIM, a especificidade enzimática define se a busca será feita em um espaço tríptico, semi-tríptico ou não-tríptico.

Com a digestão realizada nas sequências proteicas, o próximo passo é verificar quais modificações cada peptídeo poderá sofrer. Cada uma atua de forma diferente na sequência e por isso serão tratadas separadamente.

Nas modificações estáticas, se a indicação de C-terminal ou N-terminal para as sequências peptídicas estiver ativa, somente o *DeltaMass* dos peptídeos será alterado.

Já nas modificações variáveis, para cada peptídeo, são geradas novas sequências com todas as combinações possíveis a partir daquele peptídeo. A massa de

[13] *Parser*, ou analisador sintático, é o responsável por analisar uma sequência de entrada para determinar sua estrutura gramatical de acordo com um determinada gramática pré-estabelecida.

cada novo peptídeo é calculada acrescida do *DeltaMass* do peptídeo original. Por definição, os peptídeos formados a partir das modificações variáveis terão suas sequências acrescidas com a informação do *DeltaMass* entre parênteses. Se a modificação indicar um C-terminal, a sequência peptídica virá antes da informação do *DeltaMass*, caso indique um N-terminal, a sequência virá depois.

Um dicionário na memória é criado contendo todos os peptídeos com modificações variáveis e estáticas. Logo, o dicionário estará estruturado através das chaves, que serão as massas teóricas dos peptídeos em um determinado intervalo, e os valores destas chaves, que serão as listas dos peptídeos compreendidos com o valor de suas massas teóricas. Dessa forma, para encontrar um determinado peptídeo, precisa-se saber somente qual é a sua massa teórica, tornando a busca mais rápida.

4.1.4 Leitura dos espectros experimentais

Uma vez criado o banco de dados de peptídeos, precisa-se agora obter os espectros experimentais, conhecidos como *tandem mass spectrum* - *tms*, para que a busca possa ser realizada. Tais espectros são obtidos a partir da leitura dos arquivos produzidos pelo espectrômetro de massa, sejam eles arquivos do tipo MS2[14], MGF[15] ou RAW[16]. O SIM possui um *parser* que interpreta tais arquivos, que podem ser um ou vários, obtendo assim, informações importantes para o *matching* com os espectros teóricos. Informações essas, que servem para avaliar o *tms* de acordo com o *scan number*, o tipo de dissociação da molécula (CID, ETD, HCD, ECD) – *activation type*, e, principalmente, os íons precursores e a lista de íons filhos. Estes, que são os responsáveis por mostrar o quão preciso é a razão massa/carga (*m/z*) através do parâmetro "intensidade".

Conhecido todos os espectros experimentais, estes são armazenados em uma lista de objetos para que o SIM possa então começar a realizar a busca.

4.1.5 Identificação de espectros

Para cada arquivo produzido pelo espectrômetro de massa (arquivo de entrada), é montada uma lista de espectros experimentais. A partir desta lista, é feito o

[14] O formato MS2, é usado para gravar espectros MS/MS.
[15] O formato MGF, ou *Mascot Generic Format*, é um padrão muito utilizado por diversas ferramentas de busca para gravar espectros MS2, e foi oriundo da ferramenta *Mascot* [27].
[16] O formato RAW é um formato proprietário da empresa *Thermo Scientific*[39], que é gerado pelos espectrômetros de massa desta fabricante, como por exemplo, o *Orbitrap*.

confronto dos *tms* e dos espectros teóricos, resultando, assim, na quantidade de peptídeos reconhecidos para aquele arquivo de entrada.

O responsável por realizar a comparação entre os espectros no SIM é o motor de busca (*Search Engine*). É a partir dele que serão realizadas todas as buscas entre o banco de dados dos peptídeos[4.1.3] e a lista de espectros experimentais[4.1.4].

A busca é realizada caso exista para cada *tms* analisado uma quantidade mínima de picos contidos no envelope monoisotópico[17], e, também, se existir ao menos um íon precursor cuja massa MH for maior que o mínimo pré-determinando (*threshold*). Tanto a quantidade mínima de picos, quanto o *threshold* são pré-definidos nos parâmetros iniciais do SIM[4.1.1]. A partir de então, para que a *Search Engine* possa analisar os espectros experimentais, todos os íons contidos no *tms* serão normalizados a partir do íon de maior pico, ou seja, aquele que contenha a maior intensidade. Uma vez normalizados, a lista de íons será ordenada ascendentemente e os íons com intensidade menor que um valor pré-determinado serão descartados. Valor este denominado de *PeakRankThreshold*, que é definido nos parâmetros do SIM.

Cada lista de *tms* pode conter um ou mais íons precursores, os quais contém a massa MH e a carga Z; propriedades fundamentais para a busca de peptídeos no banco de dados.

Para cada precursor existente, uma lista de peptídeos candidatos será obtida. Esta lista é feita a partir dos peptídeos mais relevantes de acordo com sua massa. Esta, que deverá estar dentro do intervalo *ppm*, calculado a partir da massa MH do precursor analisado. Como a obtenção dos candidatos é feita através dos peptídeos mais relevantes de acordo com a massa, no momento da busca, peptídeos duplicados podem ser acrescidos. Logo, após preencher a lista, todos esses peptídeos duplicados serão removidos. Feito isso, uma análise mais refinada é realizada com todos os peptídeos candidatos. Análise essa que ordena toda a lista dos candidatos de acordo com o *OrbitrapPPM*[18], e exclui todos os peptídeos que tem sua massa maior que o limite pré-determinado nos parâmetros do SIM.

[17] Envelope monoisotópico é compreendido pelo conjunto de picos de um determinado íon precursor contendo as intensidades de acordo com sua massa e sua carga.
[18] *OrbitrapPPM* é uma acurácia realizada a partir de dados de alta resolução. A técnica de fragmentação normalmente utilizada é a *Higher-Energy Collisional Dissociation* - HCD.

4.1.5.1 Espectros teóricos

Uma vez obtida a lista de peptídeos candidatos, o próximo passo é percorrê-la para predizer os espectros teóricos. Estes são preditos a partir dos peptídeos contidos no banco de dados. Cada sequência peptídica então é analisada e os íons de cada uma são previstos.

Para obter os íons previstos, é necessário identificar os picos dos íons tipo *a*, *b*, *c*, *x*, *y*, *z*, a fim de obter os precursores. A obtenção dos íons tipo *b* e *y* é realizada a partir da fragmentação da sequência peptídica analisada. Os íons tipo *b* são os compostos pelos aminoácidos da sequência, lidos da esquerda para a direita. Já os íons tipo *y* são obtidos a partir da leitura inversa do espectro, ou seja, da direita para a esquerda. Entretanto, para melhor detalhar o espectro teórico, é necessário saber além dos aminoácidos, a carga, a massa e a intensidade a qual difere de um íon para o outro. Todas essas informações são necessárias para compor a predição de um íon. Por conseguinte, no caso dos íons tipo *b* e *y*, calcula-se a massa através do aminoácido correspondente, acrescendo a massa monoisotópica de um átomo de hidrogênio. Precisa-se verificar também se há perda neutra. Caso ocorra, é necessário retirar a massa da molécula sofrida, seja ela, a água ou o grupo amina, da razão *m/z* daquele íon. Os íons tipo *z* são obtidos removendo-se a massa de um grupo amina (NH_3) do valor *m/z* e acrescendo a massa monoisotópica do átomo de hidrogênio. Outrossim, insere-se a massa do grupo amina para obter os íons tipo *c*. Removendo-se um átomo de monóxido de carbono (CO) da razão *m/z*, obtém-se os íons tipo *a*, e por fim, ao inserir a massa de um CO, obtém-se os íons tipo *x*.

Para terminar a identificação de todos os picos de íons, é necessário prever os íons que tiveram perdas neutras, seja de água ou de amina. Uma vez identificados, resta agora encontrar os picos isotópicos para completar a lista dos íons previstos.

Para montar o arquivo *SQT*[19] – o qual é um dos arquivos de resposta do SIM – precisamos então gerar os parâmetros com as informações obtidas nos passos anteriores. Tais parâmetros, como *PrimaryScore*, *SecondaryScore*, *Peptide*, *PeaksMatched* e *PeaksConsidered*, são os responsáveis por indicar o quão confiável está cada resposta obtida, e eles são obtidos a partir dos íons previstos anteriormente.

[19] O formato de arquivo *SQT* é utilizado para gravar os *matches* entre os espectros MS/MS e uma base de dados de sequências peptídicas. O nome da extensão *SQT* foi escolhido como uma abreviação de um *software* de busca por espectrometria de massa, o *SEQUEST*.

Estes, por sua vez, serão os responsáveis pelo cálculo para a obtenção do *PrimaryScore* e do *SecondaryScore*.

4.1.6 Criando arquivo de saída

Uma vez obtida a lista de íons previstos, pode-se agora calcular o *PrimaryScore* e o *SecondaryScore*. O primeiro é calculado a partir do produto escalar entre os espectros teóricos, obtidos através da lista dos íons acima descrito, e os espectros experimentais, que são os íons contidos em cada *tms*. Este cálculo é penalizado pelo peso correspondente a cada região [4.1.6.1]. Todavia, para saber em qual região do espectro um determinado íon pertence[24], é preciso saber a massa do precursor e também a carga *z* do íon, passados como parâmetros no método. Assim sendo, a porcentagem de íons com a intensidade menor e maior, de acordo com cada região, pode ser obtida. O *PrimaryScore*, então, é fruto do resultado do produto escalar penalizado pela porcentagem obtida pela quantidade de picos dos íons preditos que tiveram o *matching* com os picos dos íons experimentais, com o total de íons preditos. Essa porcentagem é denominada de *percentagePeaksScore*.

Já o *SecondaryScore* é calculado a partir da porcentagem obtida do número de íons com intensidade em cada região do espectro, multiplicado pela *percentagePeaksScore*. Por fim, há o *PeaksConsidered*, que é o total de íons que foram analisados no respectivo *tms*.

Todavia, antes de gravar o arquivo de saída *SQT*, precisa-se definir o *Primary Rank* e o *Secondary Rank*, como também, a acurácia do resultado, para saber o quão confiável ele é. O *Primary Rank* é o *ranking* dado à saída atual de acordo com o número de íons precursores existentes para cada espectro experimental analisado. Outrossim, o *Secondary Rank* é o ranking dado à saída conforme o número de peptídeos candidatos encontrado para cada íon precursor. Para calcular a acurácia, foi utilizado o Δ*CN (Delta Correlation)*, que é obtido a partir da seguinte fórmula:

$$\Delta CN = \frac{X_{n-1} - X_n}{X_{n-1}}$$

Destarte, são detectados os aminoácidos que precedem e que pós cedem a sequência peptídica. Esta metodologia foi adotada para adaptar-se ao padrão do *SQT* e prover maiores informações sobre a determinada sequência. Assim sendo, os dados são gravados no arquivo de saída de acordo com a ordem determinada pelo padrão.

Ordem esta que contém para cada arquivo um cabeçalho explicando as linhas conseguintes.

4.1.6.1 Pesos ótimos das regiões do espectro

Para determinar os pesos em cada região do espectro, apresentado no início deste capítulo, foi necessário produzir uma lista de espectros bons e outra lista com espectros ruins. A partir destas duas listas, foi feita uma programação linear (PL), onde a função objetivo foi encontrar o maior *Score* a partir da soma dos pesos da região, com a restrição que essa soma não poderia passar de 1.0. A partir daí, conseguimos pesos ótimos no valor de 0.34 para a região 1 e 0.66 para a região 2.

Outro formato de saída que a ferramenta de busca também produz é o *.sim. Este formato foi criado a fim de simplificar os dados contidos no arquivo *SQT*. Nele, estão contidos apenas o *ScanNumber*, que refere-se ao número do espectro experimental percorrido; o *PeptideSequence*, o *PrimaryScore*, o *SecondaryScore*, o ΔCN e a massa teórica do peptídeo. Dessa forma, pode-se rapidamente obter os dados de forma simples e objetiva.

4.1.7 Interface gráfica

O SIM poderá ser executado em linha de comando a fim de ser suportado em ambientes clusterisados, onde o Sistema Operacional não aceite um *Graphic User Interface* (GUI). Contudo, para facilitar a usabilidade, foi desenvolvido uma interface gráfica de forma clara e objetiva, conforme mostra a Figura 17, para que o usuário tenha uma forma fácil de executar a busca na ferramenta, assim como configurar os parâmetros para uma melhor performance.

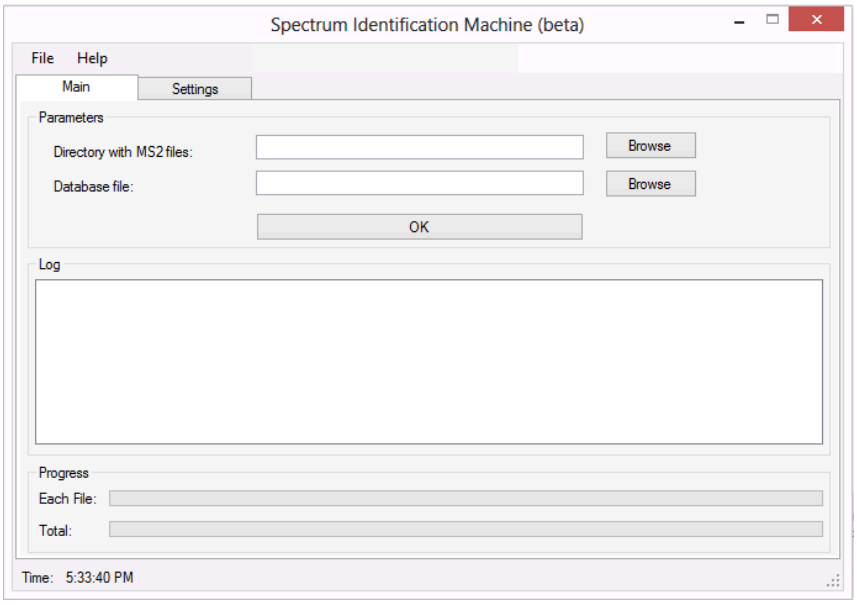


Figura 17: Interface gráfica do SIM

Ao executar o *software*, o usuário encontrará uma tela onde poderá selecionar o diretório contendo os arquivos oriundos do espectrômetro de massa com extensão *MS2, MGF ou RAW*, e também poderá selecionar o arquivo relacionado ao banco de dados, que poderá ser um arquivo FASTA[20], Target-Reverse[21] (T-R) ou Middle-Reverse/Pair-Reverse (MR-PR).

Para cada arquivo analisado, as informações referentes à busca serão apresentadas na aba *Log*, e o usuário poderá acompanhar o progresso do processo de acordo com a barra na parte inferior da interface.

Entretanto, a busca apenas poderá ser realizada após a configuração dos parâmetros do SIM, através da aba *Settings*, conforme demonstrado na Figura 18. Ali, poderão ser selecionados os valores mais apropriados para uma determinada busca, assim como a inserção e/ou remoção das modificações pós-traducionais convenientes.

Conforme explicado na seção 4.1.6, o *Spectrum Identification Machine* poderá produzir arquivos *SQT* e também arquivos em um formato próprio (sim). Também na aba *Settings*, o usuário poderá definir se deseja obter resultados com extensão *.sqt e/ou extensão *.sim.

[20] O arquivo no formato FASTA é aquele onde estão presentes sequências nucleotídicas ou sequências peptídicas, o qual os nucleotídeos ou os peptídeos estão representados por um código simples. Esta extensão é originária do software FASTA, mas hoje se tornou um padrão no ambiente proteômico.
[21] O arquivo T-R compreende aquele o qual, para cada sequência peptídica há uma sequência invertida representando o peptídeo *decoy* (peptídeo falso).

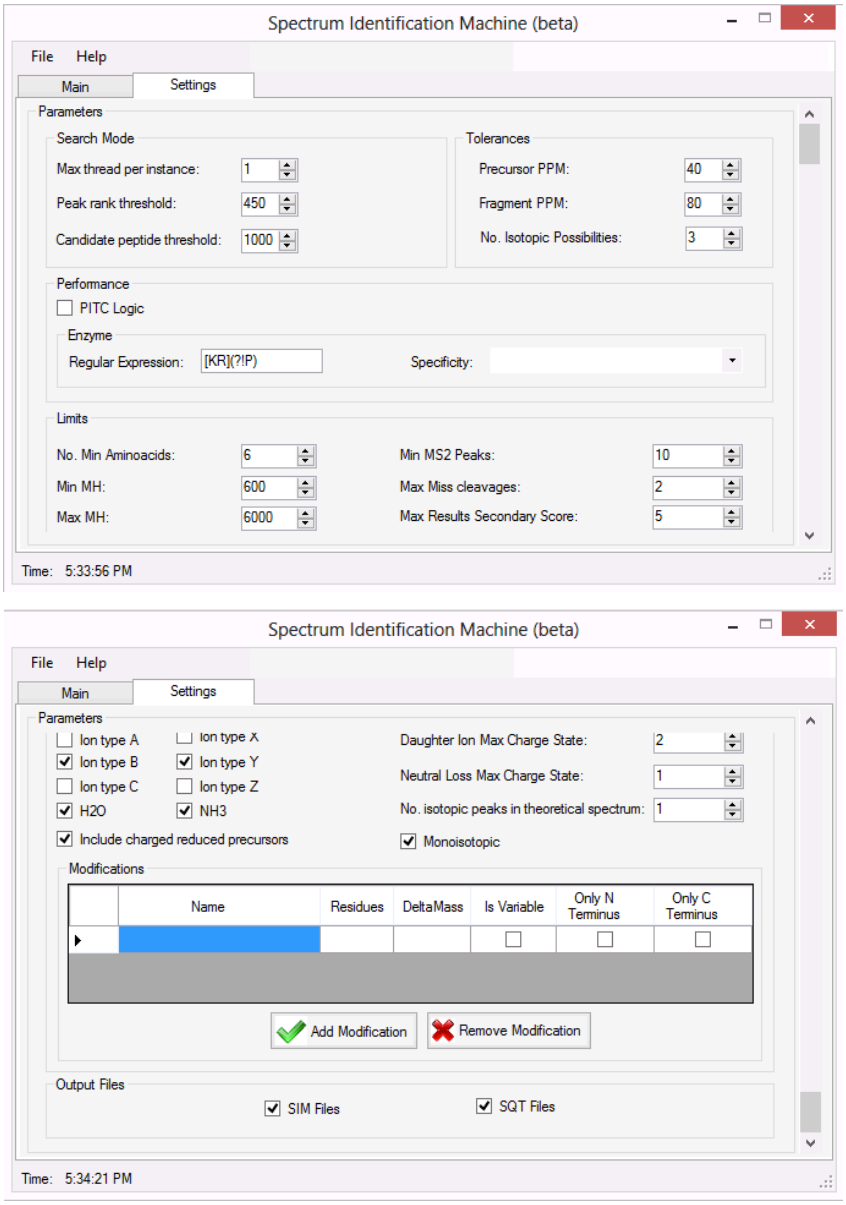


Figura 18: Aba de configurações do SIM; pode-se configurar os parâmetros mais apropriados de acordo com a busca, além de incluir determinadas modificações que a proteína poderá sofrer.

Uma vez configurado, o usuário poderá salvar os parâmetros para uma busca posterior, simplesmente indo no menu *File*, e selecionar *Save SimParams*, ou pressionando a combinação de teclas CTRL + S. O arquivo então, será salvo em um local desejado com todos os parâmetros ali contidos.

Para carregar futuramente o *simParams* – arquivo no formato XML, basta o usuário ir no menu *File*, e selecionar *Load SimParams*, ou teclando a combinação ALT + L, indicando assim o arquivo desejado. Dessa forma, o *Spectrum Identification Machine* está pronto para ser executado, esperando apenas que o botão *OK* seja pressionado.

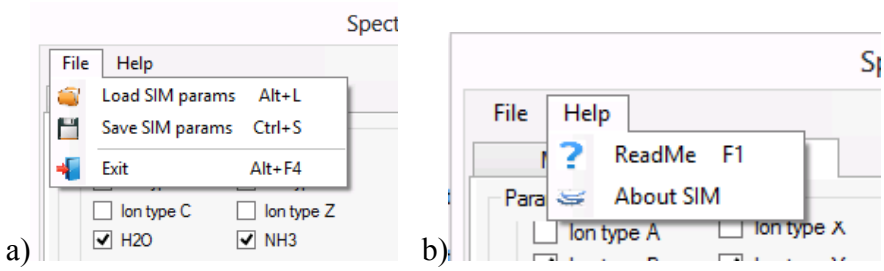


Figura 19: a – O usuário poderá carregar ou salvar o *SIM params*, evitando configurações repetitivas; b – *Ajuda* e *Sobre* do SIM

Esta ferramenta também permite a conversão de arquivos *tms*, ou seja, caso o usuário queira transformar um arquivo RAW, por exemplo, em MS2 ou MGF, basta clicar no menu *Utils*, *File Converters* e selecionar onde está o arquivo original e marcar as *checkbox's* correspondentes. Os arquivos convertidos serão gerados no mesmo diretório do arquivo original. Esta opção é útil, uma vez que, através de arquivos MS2 ou MGF pode-se visualizar os dados sem a necessidade de um *software* específico de leitura, como é o caso de arquivos do tipo RAW.

O SIM ainda possui um *Ajuda* para tirar pequenas dúvidas do usuário, assim como um *Sobre* para que o usuário possa entrar em contato conosco para esclarecer dúvidas e também para sugerir melhorias.

4.2 Lógica HI-Bone

Com o aumento do espaço de busca, a partir das modificações pós-traducionais sofridas pelas proteínas, o desempenho computacional vem a ficar debilitado. Contudo, foi desenvolvida, em conjunto com o grupo de pesquisa[22], uma estratégia experimental capaz de reduzir a complexidade do espaço de busca. Essa estratégia trata-se de um método de marcação química de peptídeos com fenil-isotiocianato, o PITC. O que esta marcação faz é aumentar a intensidade do íon tipo *b1*, correspondendo ao primeiro aminoácido da sequência peptídica, tomando como proveito a alta resolução dos espectrômetros de massa existentes hoje, que permite obter uma massa mais precisa, e também a alta intensidade dos íons b1, fazendo com que esta seja a mais intensa em uma determinada região do espectro. Como pode ser observado na Figura 20, o primeiro aminoácido da sequência peptídica

[22] O grupo de pesquisa é composto por Diogo Borges Lima, COPPE/UFRJ, Brasil; Yasset Perez-Riverol, CIGB-Cuba / EBI-UK; Aniel, CIGB-Cuba; Felipe da Veiga Leprevost, Fiocruz – PR, Brasil; Fabio C. S. Nogueira, IQ – UFRJ, Brasil; Gilberto B Domont, IQ – UFRJ, Brasil; Valmir C Barbosa, COPPE / UFRJ, Brasil; Felipe Maia Galvão França, COPPE/UFRJ, Brasil; Paulo Costa Carvalho, Fiocruz – PR, Brasil

RYPDLTLHR, representado pela arginina (R), está marcado com o fenil-isotiocianato, fazendo com que seu pico seja muito intenso, com o valor de 292.12 Da.

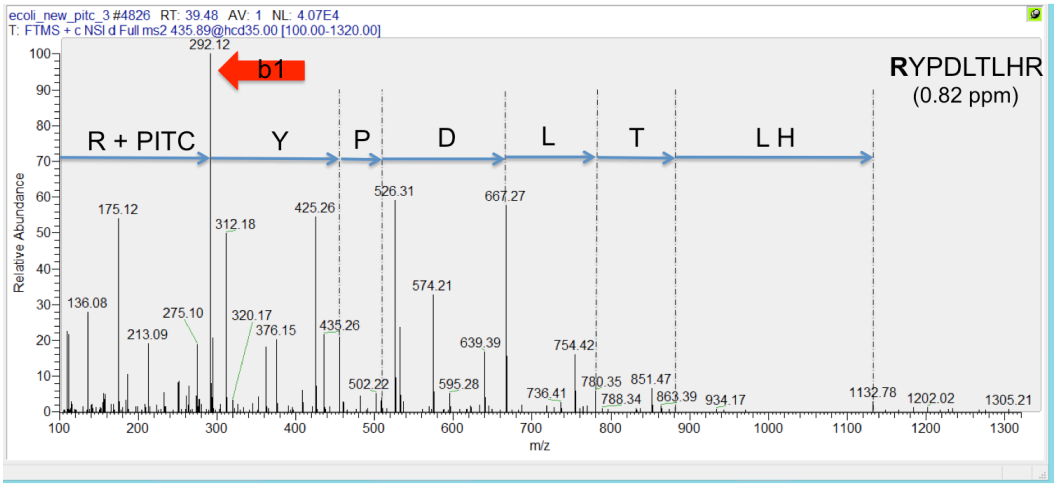


Figura 20: Exemplo de espectro cuja sequência peptídica está marcada com fenil-isotiocianato (PITC). O primeiro aminoácido, arginina, ou seja, o íon do tipo b1, é o pico mais intenso do espectro.

Foi desenvolvida então, uma lógica capaz de explorar essa intensidade. Denominada HI-Bone – ***High Intensity of b one*** *(b1)*, ela capta a intensidade do primeiro aminoácido, fazendo com que a complexidade do espaço de busca seja simplificada.

Considerando o espectro experimental da Figura 20 como exemplo e o espaço não-tríptico apresentado na Tabela 1, a seguir será demonstrado como a marcação PITC diminui a complexidade do espaço de busca demasiadamente.

Tabela 3: Exemplo do funcionamento da lógica HI-Bone

Peptídeos não marcados com PITC	Peptídeos marcados com PITC	
SPTEGLKID,	~~SPTEGLKID,~~	~~SPTEGLKID,~~
SPTEGLKI,	~~SPTEGLKI,~~	~~SPTEGLKI,~~
SPTEGLK, SPTEGL,	~~SPTEGLK, SPTEGL,~~	~~SPTEGLK, SPTEGL,~~
SPTEG, SPTE, ~~SPT,~~	~~SPTEG, SPTE, SPT,~~	~~SPTEG, SPTE, SPT,~~
~~SP, S,~~ PTEGLKID,	~~SP, S, PTEGLKID,~~	~~SP, S, PTEGLKID,~~
PTEGLKI, PTEGL,	~~PTEGLKI, PTEGL,~~	~~PTEGLKI, PTEGL,~~
PTEG, ~~PTE, PT, P,~~	~~PTEG, PTE, PT, P,~~	~~PTEG, PTE, PT, P,~~
TEGLKID, TEGLKI,	~~TEGLKID, TEGLKI,~~	~~TEGLKID, TEGLKI,~~
TEGLK, TEGL, ~~TEG,~~	~~TEGLK, TEGL, TEG,~~	~~TEGLK, TEGL, TEG,~~
~~TE, T,~~ EGLKID,	~~TE, T, EGLKID,~~	~~TE, T, EGLKID,~~
EGLKI, EGLK, ~~EGL,~~	~~EGLKI, EGLK, EGL,~~	~~EGLKI, EGLK, EGL,~~
~~EG, E,~~-GLKID, GLKI,	~~EG, E, GLKID, GLKI,~~	~~EG, E, GLKID, GLKI,~~
~~GLK, GL, G,~~ LKID,	~~GLK, GL, G, LKID,~~	~~GLK, GL, G, LKID,~~
~~LKI, LK, L, KID, KI,~~	~~LKI, LK, L, KID, KI,~~	~~LKI, LK, L, KID, KI,~~
~~K, ID, I, D, K, GLK,~~	~~K, ID, I, D, K, GLK,~~	~~K, ID, I, D, K, GLK,~~
EGLK, TEGLK,	~~EGLK, TEGLK,~~	~~EGLK, TEGLK,~~
PTEGLK,	~~PTEGLK,~~	~~PTEGLK,~~
RSPTEGLK,	**RSPTEGLK,**	**RSPTEGLK,**
RSPTEGL, RSPTEG,	**RSPTEGL, RSPTEG,**	~~**RSPTEGL, RSPTEG,**~~
RSPTE, RSPT, ~~RSP,~~	**RSPTE, RSPT,** ~~RSP,~~	~~**RSPTE, RSPT**~~**,** ~~RSP,~~
~~RS, R,~~ ARSPTEGLK,	~~RS, R, ARSPTEGLK,~~	~~RS, R, ARSPTEGLK,~~
ARSPTEGL,	~~ARSPTEGL,~~	~~ARSPTEGL,~~

ARSPTEG, ARSPTE, ARSPT, ARSP, ~~ARS~~, ~~AR~~, ~~A~~	~~ARSPTEG, ARSPTE, ARSPT, ARSP, ARS~~, ~~AR~~, ~~A~~	~~ARSPTEG, ARSPTE, ARSPT, ARSP, ARS~~, ~~AR~~, ~~A~~
34 peptídeos	**5 peptídeos**	**1 peptídeo**

Como se pode observar, na primeira coluna, tem-se o espaço de busca onde os peptídeos não foram marcados com PITC. Este, contém um número grande, totalizando 34. Contudo, o objetivo deste exemplo é procurar todas as sequências cujo primeiro aminoácido seja a arginina, representado por *R*, uma vez que, no espectro experimental, este foi o aminoácido que teve o pico mais intenso. Logo, o espaço reduziu de tamanho, chegando a cinco peptídeos, demonstrado na segunda coluna. Entretanto, os peptídeos marcados com fenil-isotiocianato e que não possuem a lisina, representada por *K*, possuem sua carga neutralizada. Isto é uma característica bioquímica da marcação. Logo, estes peptídeos também são desconsiderados no momento da busca. Por fim, chega-se a conclusão de que apenas uma sequência está apta a ser candidata do espectro experimental apresentado, como demostrado na última coluna, reduzindo assim o espaço de busca, antes de 34, para apenas um peptídeo.

4.2.1 A lógica HI-Bone e o SIM

O HI-Bone foi implementado no SIM, aumentando, significativamente a sensibilidade da ferramenta de busca. Sua implementação foi feita considerando o primeiro aminoácido da sequência peptídica do espectro experimental, assim como o segundo aminoácido, caso a intensidade do pico caísse em empate, de acordo com uma determinada margem de erro. Foram consideradas também as sequências que contivessem arginina, uma vez que este aminoácido estava presente em 92% dos peptídeos marcados com PITC, de acordo com os experimentos realizados.

Considerando a Tabela 2, a massa do PITC (135,0143 Da) foi acrescida a cada aminoácido, juntamente com a massa do hidrogênio (1,007825032), a fim de obter como resultado o pico que tenha a maior intensidade de uma determinada região do espectro. Levando em consideração que a marcação poderá ocorrer na C (cisteína), no N-terminal, ou de forma estática, de acordo com as propriedades químicas, as massas ficaram dispostas de acordo com a Tabela 4:

Tabela 4: Massa dos aminoácidos acrescida de PITC

Aminoácido / Molécula	Descrição (com PITC)	Massa
G	Glicina	193,0435887
A	Alanina	207,0592388
S	Serina	223,0541534
P	Prolina	233,0748888
V	Valina	235,0905389
T	Treonina	237,069803
C	Cisteína (acrescida a massa da modificação igual a 71,03711)	310,06842
I	Isoleucina	249,106189
L	Leucina	249,106189
N	Aspargina	250,065052
D	Ácido Aspártico	251,049068
Q	Glutamina	264,080703
K	Lisina	264,117088
E	Ácido Glutâmico	265,064718
M	Metionina	267,06261
H	Histidina	273,081037
F	Fenilanina	283,090539
U	Selenocisteína	286,975765
R	Arginina	292,123236
X ou J	Leucina ou Isoleucina	249,106185
Y	Tirosina	299,085454
W	Triptofan	322,101438
O	Pirrolisina - O 22º aminoácido	391,188817
H	Hidrogênio	137,0299501
O	Oxigênio	152,0170397
C	Carbono	148,022125
N	Nitrogênio	150,025199

5 Resultados

Para demonstrar a robustez da ferramenta de busca desenvolvida foi necessário confrontá-la com outra já existente. Obtemos então um banco de dados de *P. Furiosos*, uma espécie de padrão de benchmark de ferramentas de busca, e rodamos o SIM sob esse banco. Confrontamos então, com o *software* Andromeda pertencente ao ambiente computacional MaxQuant, pedindo para que um usuário experiente nessa ferramenta pudesse rodar esse mesmo banco de dados, utilizando as mesmas restrições aplicadas ao SIM. O resultado obtido com um FDR de 1% está apresentado na Figura 21. Nela pode ser comprovada a robustez do *Spectrum Identification Machine*.

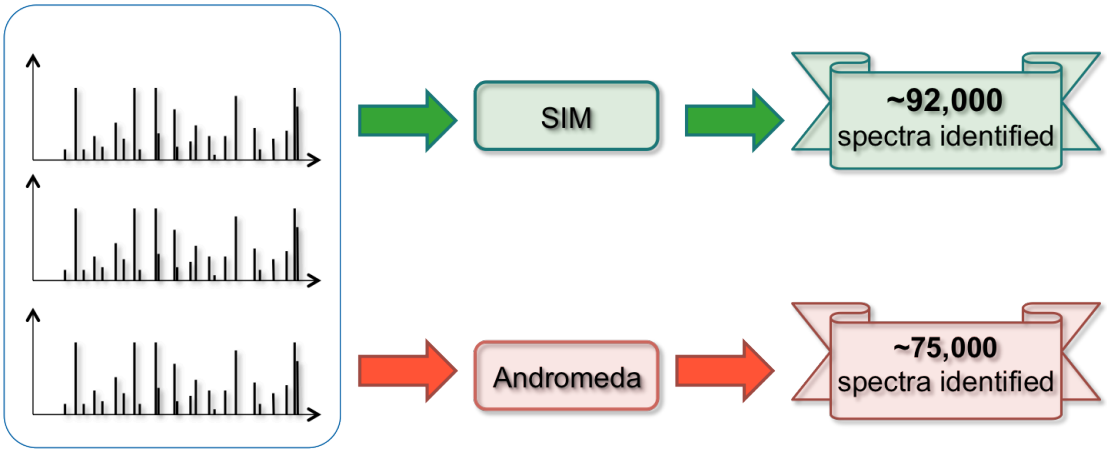


Figura 21: Comparação da sensibilidade das ferramentas ao utilizar um banco de dados de *P. Furiosos* desenvolvido no Laboratório do *John Yates III* para *benchmark*. Enquanto o SIM identificou cerca de 92 mil espectros, o Andromeda identificou 75 mil, comprovando a robustez da ferramenta desenvolvida.

O próximo passo então é realizar a busca ativando a lógica HI-Bone[4.2]. Evidentemente, o objetivo não é comparar ferramentas de buscas, uma vez que, hoje, nenhuma tem implementada a lógica HI-Bone ou semelhante a ela, somente o SIM.

O resultado obtido utilizando o banco de dados da *E.coli*., é demonstrado na Figura 22. No espaço tríptico, o menor de todos, a eficiência da nossa ferramenta mostra o quão sensível é a busca quando a lógica HI-Bone está habilitada. Nesse espaço, foram encontrados 2.412 espectros com a lógica ativa, o que representa um aumento de 16% em relação à busca realizada quando a lógica está desabilitada. Da mesma forma acontece quando vamos para o espaço semi-tríptico, porém o número de identificações neste espaço é maior [1.6.1]. E finalmente, no espaço não tríptico,

que é um espaço complexo, cujo tamanho é maior que cinquenta milhões de peptídeos[1.6.1], fica claro o ganho do resultado quando a lógica HI-Bone está ativa. Consegue-se encontrar 2.952 espectros contra 1.352 com a lógica desativada, um aumento de 118%.

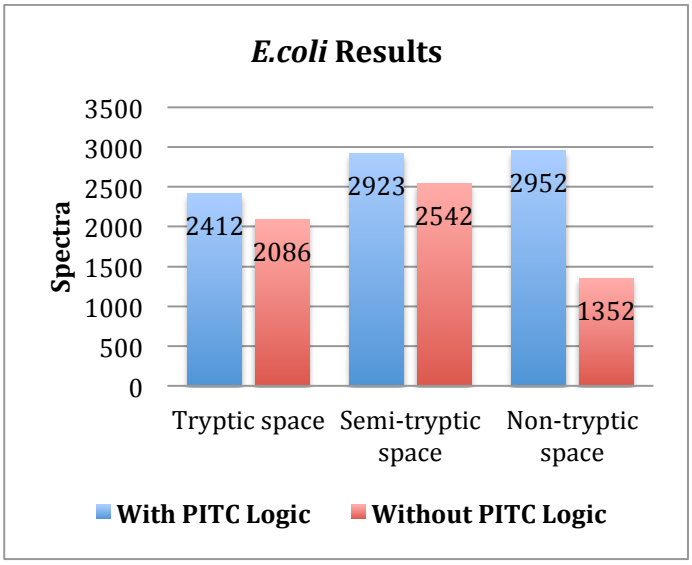


Figura 22: Resultados obtidos utilizando o banco de dados da *E.coli*, ativando ou não a lógica que explora da intensidade do íon *b1*. No espaço não-tríptico, cujo o tamanho é maior que cinquenta milhões de sequências peptídicas, obteve-se o maior número de identificação, mostrando a eficácia da lógica.

Em geral, se a complexidade do espaço de busca está sendo reduzida, o tempo de processamento também decai. Na Figura 23, é mostrada a diferença de tempo realizando a busca com e sem a lógica HI-Bone. Entretanto, no espaço mais complexo houve um pequeno aumento no tempo, devido ao número de candidatos que foram previamente excluídos. Tais candidatos são os representados no exemplo da Tabela 3, onde no espaço não tríptico o número de exclusão aumenta.

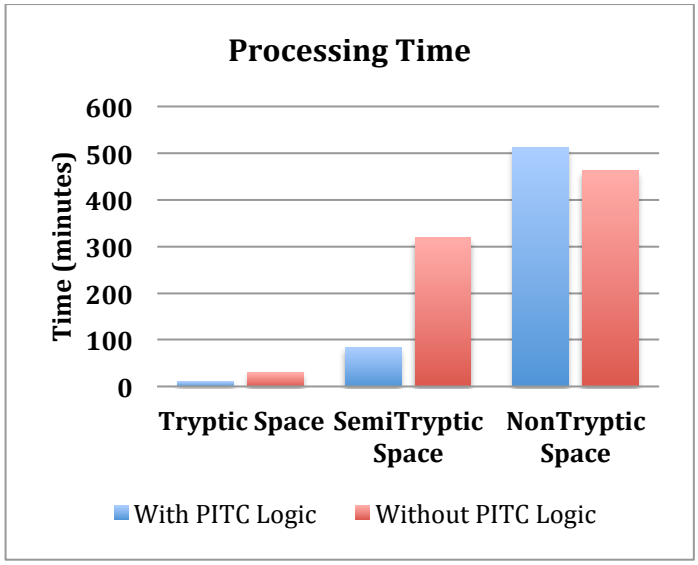


Figura 23: Tempo de processamento dos espaços de busca utilizando ou não a lógica HI-Bone.

6 Discussão e Conclusões

O *Spectrum Identification Machine* – SIM, é uma ferramenta de busca desenvolvida com base na metodologia PSM – *Peptide Spectrum Match*, o qual demonstrou uma superioridade no desempenho, através do número de espectros identificados, em relação a uma outra ferramenta pertencente ao ambiente computacional MaxQuant, o Andromeda.

A Lógica HI-Bone, implementada no SIM, capaz de explorar a intensidade do íon *b1*, correspondente ao pico mais intenso de uma determinada região do espectro, aumentou exorbitantemente o número de identificações de espectros em espaços de busca complexos, compreendidos com mais de cinquenta milhões de peptídeos. Este aumento representou em mais de 100% em relação às buscas realizadas sem a utilização da lógica, confirmando a eficácia da sensibilidade da ferramenta de busca desenvolvida.

O SIM está integrado ao *PatternLab*[13], um ambiente computacional, que contém ferramentas para realizar análises de proteômica quantitativa, análises de proteínas diferencialmente expressas, produção de Diagramas de *Venn* para a disponibilidade proteica em uma amostra, análises do *Gene Ontology*, entre outros *softwares* relacionados à proteômica computacional.

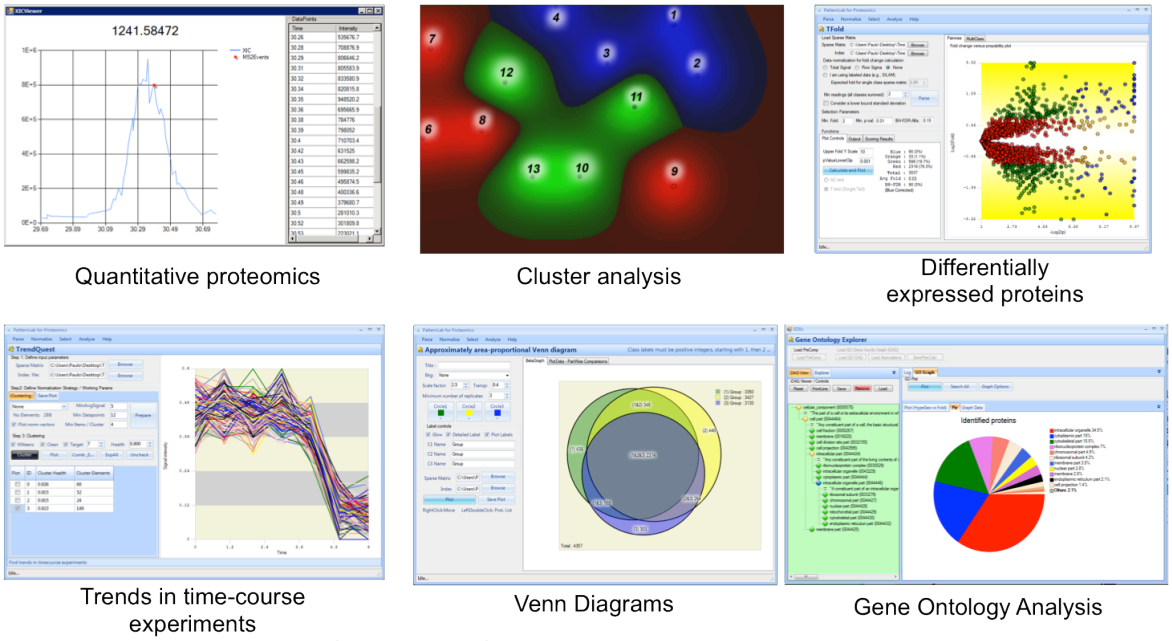


Figura 24: PatternLab – Ambiente de proteômica computacional, onde contém várias ferramentas de proteômica quantitativa, análises de *clusters*, diagrama de Venn etc.

Como perspectivas, o SIM foi integrado a uma nova ferramenta desenvolvida pelo Felipe Leprevost *et al.*, denominada *PepExplorer*, capaz de fazer análises utilizando a metodologia *de novo sequencing*. A ideia é integrar uma ferramenta que utiliza o *peptide spectrum match* com o *de novo sequencing* em um mesmo *pipeline*, gerando assim, um relatório final unificado.

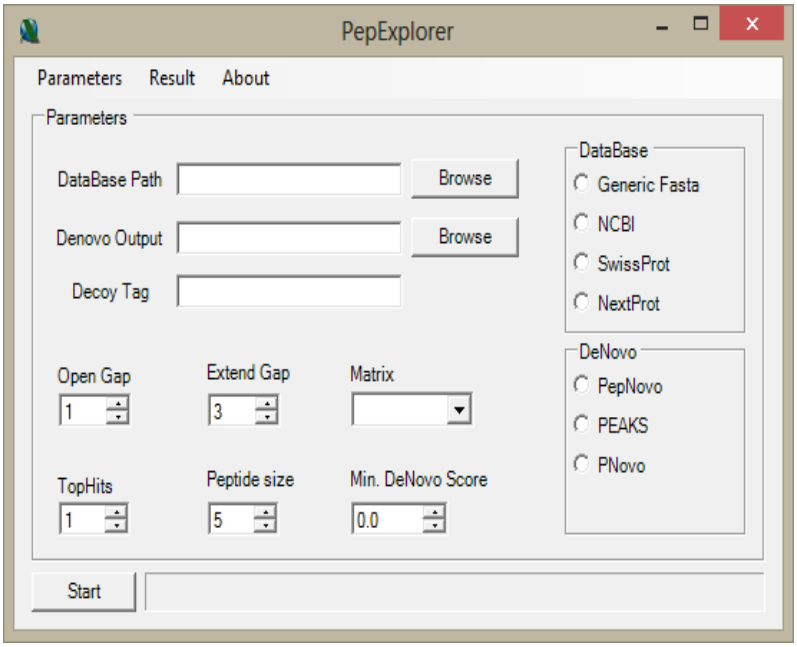


Figura 25: PepExplorer – Uma ferramenta computacional capaz de realizar análises utilizando o *de novo sequencing*.

A Figura 26 demonstra graficamente o *pipeline* integrando o *peptide spectrum match* ao *de novo sequencing* gerando um único relatório final. Dessa forma consegue realizar um estudo proteômico completo, abrangendo o sequenciamento de um peptídeo quando a proteína está contida em um banco de dados, como também fazer a análise de peptídeos ainda não identificados.

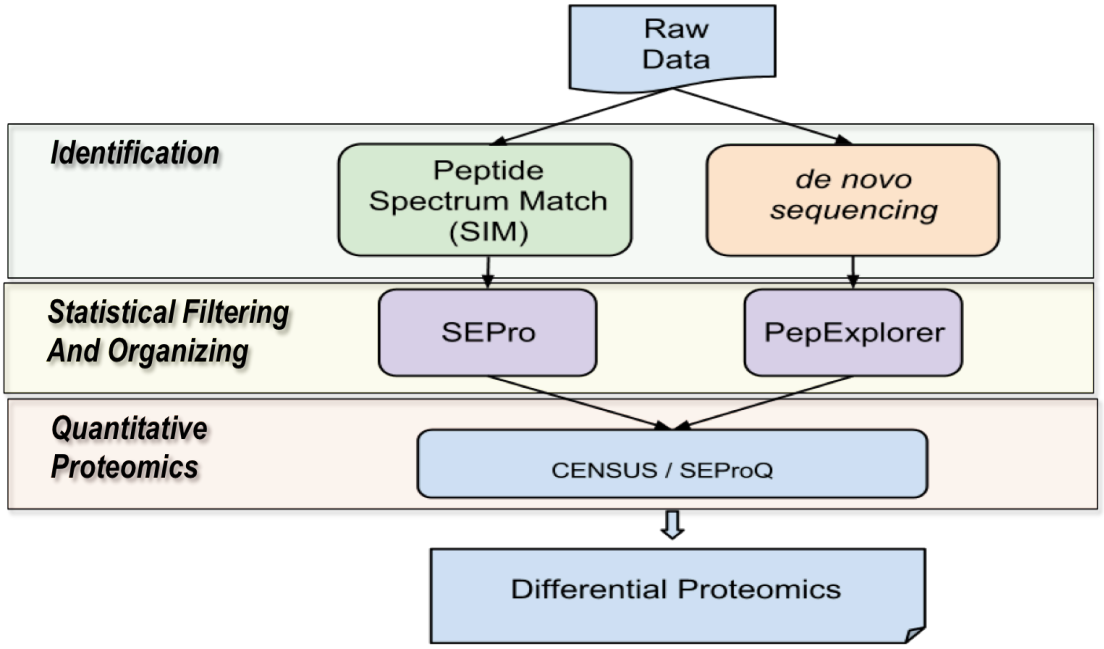


Figura 26: Pipeline integrando o *peptide spectrum match* – PSM com o *de novo sequencing*. A ideia é combinar as duas técnicas para obter um relatório final unificado.

7 Bibliografia

[1] Frank, A., & Pevzner, P. (13 de Janeiro de 2005). PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Analytical Chemistry* , 964-973.

[2] Unimod. (s.d.). Acesso em 6 de Janeiro de 2013, disponível em http://www.unimod.org

[3] Walson, J., Baker, T., Bell, S., Gann, A., Levine, M., & Losick, R. (2004). *Molecular Biology of the Gene* (5ª ed.). San Francisco, CA: Pearson, Benjamin Cummings.

[4] *Yates,J.R., III et al. (2012) Toward objective evaluation of proteomic algorithms. Nat.Methods,9, 455-456. .*

[5] *Yen,C.Y. et al. (2006) Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. Anal.Chem.,78, 1071-1084.*

[6] Aeborsold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature , 422*, 198-207.

[7] Alberts, B., Johnson, A., Lewis, J., Kazuo, Ralf, M., Roberts, K., et al. (2002). *Molecular Biology of the Cell* (4 ed.). Nova Iorque e Londes: Garland Science.

[8] Bianconi, M. L. (s.d.). *Tripsina e Quimiotripsina*. (IBqM/UFRJ) Acesso em 17 de Novembro de 2012, disponível em http://www2.bioqmed.ufrj.br/enzimas/proteases2.htm

[9] *Bioquímica: Proteína*. (s.d.). Acesso em 26 de Dezembro de 2012, disponível em http://desenvolvimentovirtual.com/bioq/InfOnline1/3%20-%20aminoacido_proteina/slides/aula_6_pt.pdf

[10] Boghigian, B. (29 de Julho de 2005). *Advances in analytical biochemistry and systems biology: Proteomics*. Acesso em 26 de Dezembro de 2012, disponível em Advances in analytical biochemistry and systems biology: Proteomics: http://openwetware.org/images/f/f2/TopicSeminarProteomics.pdf

[11] Carvalho, P. C. (Março 2010). *Um Ambiente Computacional para a Proteômica.* UFRJ, PESC - COPPE, Rio de Janeiro.

[12] Carvalho, P. C., Fischer, J. S., Xu, T., Cociorva, D., Balbuena, T. S., Valente, R. H., et al. (2012). Search engine processor: Filtering and organizing peptide spectrum matches. *Proteomics* , 944-949.

[13] *Carvalho,P.C. et al. (2008) PatternLab for proteomics: a tool for differential shotgun proteomics. BMC.Bioinformatics.,9, 316-.*

[14] *Carvalho,P.C., Yates, III., Jr., Barbosa, V. C. (2010) Analyzing shotgun proteomic data with PatternLab for proteomics. Curr.Protoc.Bioinformatics.,Chapter 13, Unit-15.* .

[15] Chait, B. (2006). Chemistry. Mass spectrometry: bottom-up or top-down? *Science , 314*, 65.

[16] Chi H, C. H. (28 de Dezembro de 2012). pNovo+: De Novo Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra. *Journal of Proteome Research* .

[17] *Cox,J. et al. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J.Proteome.Res.,10, 1794-1805.* .

[18] Cociorva, D., Tabb, L., & Yates, J. (2007). Validation of tandem mass spectrometry database search results using DTASelect. In: *Curr Protoc Bioinformatics.*

[19] *Eng,J.K. et al. (1994) An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J Am Soc Mass Spectrom,5, 976-989.* .

[20] FAPERJ. (1 de Janeiro de 2005). *Avanços nas Redes Genômica e Proteômica*. (FAPERJ, Produtor) Acesso em 28 de Dezembro de 2012, disponível em http://www.faperj.br/interna.phtml?obj_id=1953

[21] *Human Proteome Project (HPP)*. (2010). Acesso em 6 de Janeiro de 2013, disponível em http://www.hupo.org/research/hpp/

[22] *http://www.genome.gov/12011238*. (13 de Outubro de 2011). Acesso em 3 de Outubro de 2010

[23] *http://fields.scripps.edu/sequest/SQTFormat.html*. (25 de Junho de 2002). Acesso em 3 de Novembro de 2012

[24] Junqueira, M., Spirin, V., Balbuena, T. S., Waridel, P., Surendranath, V., Kryukov, G., et al. (Agosto de 2008). Separating the Wheat from the Chaff: Unbiased Filtering of background tandem mass spectra improves protein identification. *Journal of Proteome Research* , 3382-3395.

[25] Jaeger, K. E., & Eggert, T. (2004). Enantioselective biocatalysis optimized by directed evolution. In: *Curr Opin Biotechnol* (pp. 305-313).

[26] James, P. (1997). Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly reviews of biophysics , 30*, 279-331.

[27] Koenig, T., Menze, B., & et al. (Setembro de 2008). Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *Journal Proteome Research* , 3708-17.

[28] Lima, D. B. (Março de 2010). *Redução do acoplamento com frameworks específicos de plataforma no MDARTE: Estudo de caso em*

ambientes móveis. Acesso em 26 de Dezembro de 2012, disponível em Redução do acoplamento com frameworks específicos de plataforma no MDARTE: Estudo de caso em ambientes móveis: http://www.cos.ufrj.br/~diogobor/files/projetoFinal.pdf

[29] Murray, P. (2004). *Microbiologia Médica* (Vol. 4ª). Elsevier.

[30] Marques de Sá, J. (2000). *Reconhecimento de Padrões*. Acesso em 6 de Janeiro de 2013, disponível em http://paginas.fe.up.pt/~jmsa/recpad/index.htm

[31] *Matrix Science*. (2012). Acesso em 04 de Janeiro de 2013, disponível em http://www.matrixscience.com/help/percolator_help.html

[32] Olsen, J. V., Ong, S.-E., & Mann, M. (2004). Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Molecular & Cellular Proteomics , 3*, 608-14.

[33] *Peaks - Complete Software for Proteomics*. (s.d.). Acesso em 6 de Janeiro de 2013, disponível em http://www.bioinfor.com/peaks/features/overview.html

[34] Perkins, D., Pappin, D., Creasy, D., & Cottrell, J. (December, 1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. In: *Electrophoresis* (pp. 3551-67).

[35] Salvato, F., & Labate, C. A. (2007). *Modificações pós-traducionais de proteínas.* Universidade de São Paulo, Departamento de Genética, São Paulo.

[36] Sanchez, A., Perez-Riverol, Y., González, L. J., Noda, J., Betancourt, L., Ramos, Y., et al. (2010). *Evaluation of phenylthiocarbamoyl-derivatized peptides by electrospray ionization mass spectrometry: selective isolation and analysis of modified multiply charged peptides for liquid chromatography-tandem mass spectrometry experiments.* Center for Genetic Engineering and Biotechnology, Proteomics Department. Havana: NCBI.

[37] *SEQUEST*. (s.d.). (U. -A. Proteomics, Produtor) Acesso em 27 de Dezembro de 2012, disponível em http://proteomicsresource.washington.edu/sequest.php

[38] *Teoria da Probabilidade*. (s.d.). Acesso em 13 de Novembro de 2012, disponível em http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0210463_06_cap_04.pdf

[39] *Thermo Fisher Scientific Inc.* (2012). Acesso em 15 de Novembro de 2012, disponível em http://www.thermoscientific.com/ecomm/servlet/productsdetail_11152_L10727_87170_13901130_-1?ca=orbitrapelite

Anexo I

Application Note

Effectively addressing complex proteomic search spaces with peptide spectrum matching

Diogo Borges[1,*], Yasset Perez-Riverol[2,3,*], Fabio C S Nogueira[4], Gilberto B Domont[4], Jesus Noda[2], Felipe da Veiga Leprevost[5], Vladimir Besada[2], Felipe M G França[1], Valmir C Barbosa[1], Aniel Sánchez[2] & Paulo C Carvalho[5]

[1] Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[2] Proteomics Department, Center for Genetic Engineering and Biotechnology, Cubanacán, Playa, Ciudad de la Habana, Cuba

[3] Proteomic Services, EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

[4] Proteomics Unit, Institute of Chemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[5] Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil

*Equal contributions

Associate Editor: Prof. Martin Bishop

Summary: Protein identification by mass spectrometry is commonly accomplished using a peptide sequence matching (PSM) search algorithm, whose sensitivity varies inversely with the size of the sequence database and the number of post-translational modifications considered. We present the Spectrum Identification Machine, a PSM tool that capitalizes on the high-intensity b1-fragment ion of tandem mass spectra of peptides coupled in solution with phenylisotiocyanate to confidently sequence the first amino acid and ultimately reduce the search space. We demonstrate that in complex search spaces a gain of some 120% in sensitivity can be achieved.

Availability: All data generated and the software are freely available for academic use at http://proteomics.fiocruz.br/software/sim.

Contact: paulo@pcarvalho.com

1 INTRODUCTION

One of the goals of shotgun proteomics is to perform large-scale identification and quantitation of thousands of proteins within complex protein mixtures (e.g., biological fluids or whole-cell lysates). The strategy comprises protein digestion, followed by peptide chromatographic separation online with tandem mass spectrometry (MS2) (Washburn, M.P., Wolters, D., and Yates, J.R., III,2001). The MS2 data are then generally identified using a peptide sequence matching (PSM) tool; examples are SEQUEST (Eng, J.K. et al.,1994), and most recently, Andromeda (Cox, J. et al.,2011). Briefly, given a peptide's precursor ion mass and MS2, these algorithms pull out, from a peptide-sequence database, peptide sequences whose theoretical mass lies within a given tolerance from the experimental precursor mass. Following that, theoretical spectra are generated for all peptide candidates so that some similarity metric, be it empirical or statistical, can be used to select the most likely candidate. Finally, a list of identifications satisfying some false-discovery rate (FDR) is obtained by using a statistical filtering tool such as SEPro (Carvalho, P.C. et al.,2012).

The sensitivity of a PSM tool varies inversely with the size of the sequence database and the number of post-translational modifications considered (Yen, C.Y. et al.,2006). Consequently, studies addressing complex search spaces are challenging when seen from a computational perspective. Examples are analyzing snake venoms for identifying naturally occurring peptides (Tashima, A.K. et al.,2012), or performing a meta-proteomic study of a micro-organism biota (Muth, T. et al.,2012). The former requires not trypsinizing the samples and thus lifts the constraints of a PSM search engine to only tryptic peptides, which results in an exponential growth of the search space; the latter entails the concatenation of hundreds of sequence databases of different organisms. Nevertheless, the rewards at stake could be discovering a naturally occurring peptide with pharmaceutical properties or the in-depth comprehension of a system's biology.

Recently, Sánchez & Perez-Riverol et al. demonstrated the possibility to identify peptides using the N-terminal residue and accurate precursor mass; for this, they coupled peptides in solution with phenylisotiocyanate (PITC) (Sanchez, A. et al.,2010). During the activation in the collision cell, these phenylthiocarbamoyl-derivatized peptides dissociate to specifically yield an intense b1 fragment. This unlocks the possibility to confidently determine the N-terminal residue in a single mass spectrum. The authors then demonstrated a peptide identification tool that considered only the b1 fragment ion mass and the high mass accuracy of the precursor, and used it to identify peptides in an *Escherichia coli* tryptic digest. The shortcomings of this method are in the inability to discriminate between peptides with close masses and same first residue. As the remaining MS2 information is not taken into account, the method is blind to peptides not found in the database but also coinciding in mass and first residue, and thus prone to such false positives. More on these limitations is found in a discussion in the supplementary file. That said, this strategy becomes inapplicable to studies addressing complex search

spaces, where these "coincidences" become increasingly frequent. Notwithstanding this, the authors demonstrated a way to potentially improve current PSM algorithms.

2 METHODS

To overcome these limitations, we present the Spectrum Identification Machine (SIM). SIM capitalizes on PITC-coupled peptides to reduce the search space by filtering peptide candidates to only those satisfying the precursor mass and the first amino acid obtained from the high-intensity b1 fragment. The reduced search space is then queried by comparing theoretically generated spectra to experimental ones with a similarity metric that is the dot-product between the normalized experimental and theoretical spectra, multiplied by the number of matched peaks. This enables the selection of the highest-scoring candidate sequence. Some other scores, such as DeltaCN from SEQUEST, are also computed; in fact, the output of SIM is a .SQT file (i.e., it has the SEQUEST output format), which makes every tool that works with SEQUEST automatically compatible with SIM.

We benchmarked SIM, with results filtered by SEPro to achieve a 1% FDR (protein level), on a previously published yeast lysate MudPIT dataset (Barboza, R. et al.,2011) against the widely adopted Andromeda. We note that this is a non-PITC-labeled dataset, so this benchmarking was carried out to verify whether SIM would perform acceptably. Search parameters and results are available at the SIM website. In our hands, Andromeda (v. 1.3.0.5) identified 53,997 MS/MS and SIM (v 0.905) 73,639 MS/MS, both constrained by the same FDR of 1% at the protein level. This result demonstrates that SIM does indeed have an effective algorithm for PSM and has allowed us to focus our efforts on showing the benefits of activating what we term the PITC logic.

We verify the efficiency of the PITC logic on a PITC-labeled *E. coli* extract that was trypsinized and analyzed with a one-hour reversed-phase chromatography gradient on an Orbitrap Velos acquiring MS2 in HCD mode. To verify how the increase in database complexity affected the results, we generated three peptide databases, one comprehending only fully tryptic peptides (one missed-cleavage accepted and no PTMs), the second having a semi-tryptic specificity, and the third with no enzymatic specificity. This generated search spaces comprising 566,070, 11,217,794, and 63,102,231 peptides, respectively. Results were filtered with SEPro to converge to a list of 1% FDR.

3 RESULTS

Search results with and without the PITC logic are presented in Figure 1. An example of a PITC peptide tandem mass spectrum is found in Supplementary Figure 1.

4 DISCUSSION AND CONCLUSIONS

We have searched an *E. coli* tryptic digest labeled with PITC using SIM. We performed a proof of concept by testing the efficiency of our new PITC logic under increasing complexities, i.e., from tryptic to semi-tryptic to fully tryptic, and obtained an increase in sensitivity of some 120% in a large search space. As such, the SIM-PITC approach is recommended when addressing proteomic studies with complex search spaces. SIM has a graphical user interface to provide a user-friendly experience, is multiplatform, and can be executed in cluster environments. SIM is integrated into PatternLab for proteomics (Carvalho, P.C. et al.,2008; Carvalho, P.C., Yates, I., Jr., and Barbosa, V.C.,2010), which makes available an arsenal of tools for quantitative and differential proteomics.

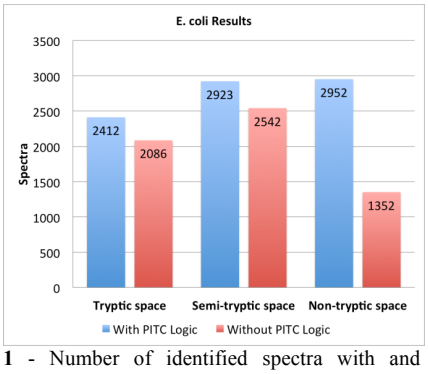


Figure 1 - Number of identified spectra with and without activating SIM's PITC logic.

5 AKNOWLEDGEMENTS

D.B. and Y.P.-R. have contributed equally to this work. The authors thank Dr. Fabrico Marchini and Michel Batista for technical discussions, and FAPERJ, CNPq, and PDTIS for financial support.

REFERENCES

Barboza,R. et al. (2011) Can the false-discovery rate be misleading? *Proteomics.*,11, 4105-4108.

Carvalho,P.C. et al. (2008) PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC.Bioinformatics.*,9, 316-

Carvalho,P.C. et al. (2012) Search engine processor: Filtering and organizing peptide spectrum matches. *Proteomics.*,12, 944-949.

Carvalho,P.C., Yates, I., Jr., Barbosa, V. C. (2010) Analyzing shotgun proteomic data with PatternLab for proteomics. *Curr.Protoc.Bioinformatics.*,Chapter 13, Unit-15.

Cox,J. et al. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J.Proteome.Res.*,10, 1794-1805.

Eng,J.K. et al. (1994) An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J Am Soc Mass Spectrom*,5, 976-989.

Muth,T. et al. (2012) Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Mol.Biosyst.*,

Perez-Riverol,Y. et al. (2011) In silico analysis of accurate proteomics, complemented by selective isolation of peptides. *J.Proteomics.*,74, 2071-2082.

Sanchez,A. et al. (2010) Evaluation of phenylthiocarbamoyl-derivatized peptides by electrospray ionization mass spectrometry: selective isolation and analysis of modified multiply charged peptides for liquid chromatography-tandem mass spectrometry experiments. *Anal.Chem.*,82, 8492-8501.

Tashima,A.K. et al. (2012) Peptidomics of three Bothrops snake venoms: insights into the molecular diversification of proteomes and peptidomes. *Mol.Cell Proteomics.*,11, 1245-1262.

Washburn,M.P., Wolters, D., Yates, J. R., III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat.Biotechnol.*,19, 242-247.

Yen,C.Y. et al. (2006) Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. *Anal.Chem.*,78, 1071-1084.