MÉTODO SEMI-ROTULADO PARA APONTAR IDENTIFICAÇÕES CONFIÁVEIS EM EXPERIMENTOS DE PROTEÔMICA *SHOTGUN*

Rodrigo de Moura Barboza

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
Junho de 2011

MÉTODO SEMI-ROTULADO PARA APONTAR IDENTIFICAÇÕES CONFIÁVEIS EM EXPERIMENTOS DE PROTEÔMICA *SHOTGUN*

Rodrigo de Moura Barboza

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.

_____
Prof. Luís Alfredo Vidal de Carvalho, D.Sc.

_____
Dra. Juliana de Saldanha da Gama Fischer Carvalho, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2011

Barboza, Rodrigo de Moura

Método semi-rotulado para apontar identificações confiáveis em experimentos de proteômica *shotgun* / Rodrigo de Moura Barboza. – Rio de Janeiro: UFRJ/COPPE 2011.

VIII, 41 p.: il.; 29,7 cm.

Orientadores: Felipe Maia Galvão França

Paulo Costa Carvalho

Dissertação (Mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2011.

Referências Bibliográficas: p.38 – 40.

1. Proteômica. 2. Reconhecimento de Padrões. 3. Espectrometria de massa. I. França, Felipe Maia Galvão *et al*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Dedico esta dissertação aos meus pais:

Tiudorico Leite Barboza e Maria Isabel de Moura Barboza

Agradecimentos

Agradeço aos meus pais, Tiudorico e Maria Isabel, que me apoiaram em todos os momentos desta jornada e são os responsáveis por tudo que já conquistei nesta vida. Sem eles nada disso seria possível.

Agradeço ao meu irmão, Carlos Alberto, que não me deixou perder o foco no momento mais difícil.

Agradeço à Raquel Castiglione, que me deu força em todos os momentos e contribuiu com sua experiência acadêmica para me auxiliar.

Agradeço a toda a minha família que sempre esteve presente com palavras de carinho e apoio.

Agradeço ao Professor Felipe M. G. França que me recebeu de braços abertos como seu aluno e me ajudou a investir meus estudos em algo que realmente me deu gosto de trabalhar.

Agradeço ao Dr. Paulo Carvalho que trabalhou junto comigo na obtenção dos resultados e mostrou-se um excelente orientador.

Agradeço ao professor João Carlos que, mesmo após a minha graduação, continuou me auxiliando com seu grande conhecimento.

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

MÉTODO SEMI-ROTULADO PARA APONTAR IDENTIFICAÇÕES CONFIÁVEIS EM EXPERIMENTOS DE PROTEÔMICA *SHOTGUN*

Rodrigo de Moura Barboza

Junho/2011

Orientadores: Felipe Maia Galvão França
Paulo Costa Carvalho

Programa: Engenharia de Sistemas e Computação

Em proteômica, o método padrão ouro para apontar identificações protéicas confiáveis por espectrometria de massa utiliza uma ferramenta de identificação (e.g., SEQUEST, ProLuCID) juntamente com uma base de dados de sequências de proteínas reais (*target)* e inexistentes (*decoy).* As sequencias *decoy* visam modelar um discriminador a partir de *scores* obtidos da uma ferramenta de busca. Em seguida, as identificações são ordenadas de acordo com uma distância da superfície de decisão do discriminador e, em geral, é aceita uma fração de 1% de identificações falsas (*decoys*) no resultado final. Neste cenário, acredita-se que outros 1% de espectros de massa identificados como *target* são falso-positivos. Neste trabalho, primeiramente mostramos uma falha no método padrão ouro: o resultado real pode não ser condizente com o erro predito; em seguida mostramos como determinar quando esta limitação ocorre e como remediá-la.

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

A SEMI-LABELED APROACH FOR DETECTING RELIABLE IDENTIFICATIONS IN SHOTGUN PROTEOMICS EXPERIMENTS

Rodrigo de Moura Barboza

June/2011

Advisors: Felipe Maia Galvão França
Paulo Costa Carvalho

Department: System and Computing Engineering

In proteomics, the gold standard method to point reliable proteins identifications by mass spectrometry uses an identification tool (e.g., SEQUEST, ProLuCID) together with a database of real protein sequences (target) and sequences that don‘t exist (decoy). The decoy is used to model a discriminator using the scores given by the search tool. After that, the identifications are ordered according to a distance from discriminator decision surface and, generally, a fraction of 1% of false identifications (decoys) is accepted in the final result. In this scenario, it is believed that other 1% of mass spectra identified as target are false positives. In this work we first show a flaw in the gold standard method: the final result may be not consistent with the predicted error; then we show how to identify when this limitation occurs and how to overcome this.

ÍNDICE

1 - Introdução

1.1 – Motivação

Um ramo importante das ciências da vida é a caracterização de sistemas biológicos pelo conjunto de proteínas que um determinado organismo está expressando no tempo e no espaço. Este campo é conhecido como *Proteômica* [1].

Uma lagarta e uma borboleta, por exemplo, possuem diferenças muito acentuadas. Uma voa, a outra rasteja, possuem formas diferentes, etc. Embora tenham proteomas muito diferentes, o seu genoma é o mesmo, ou seja, em cada uma existem proteínas que caracterizam o estado em que elas se encontram.

Figura 1 – Ilustração da morfogênese de um lepidóptero. A lagarta, à esquerda na figura, após sofrer mudanças na fase de pupa, transforma-se em uma borboleta. Figuras modificadas de [2,3].

A proteômica é bastante utilizada no estudo de patologias. Compreender diferenças entre um estado celular sadio e um doente pode ajudar a produzir novos fármacos na busca pela cura de doenças e prover conhecimento, ao nível molecular, da

patologia em questão. Para isso, faz-se necessária a identificação de proteínas características de cada estado.

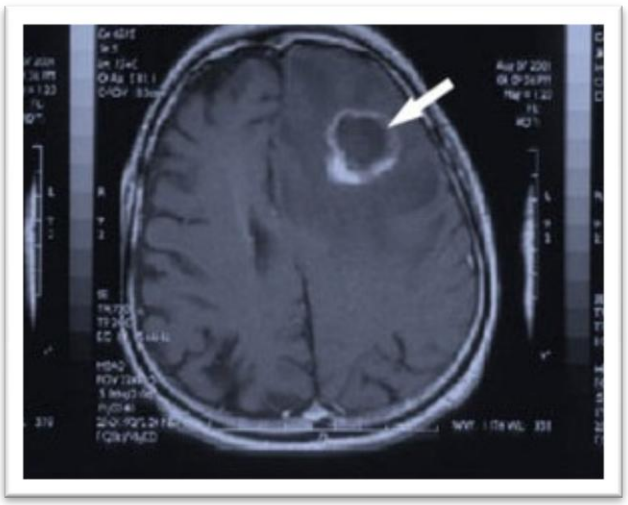

Figura 2 - Ressonância magnética de um cérebro com tumor. Figura obtida de [4].

A Figura 2 mostra a imagem de um cérebro com um tumor. A identificação de proteínas características do tumor em relação às das células sadias do cérebro, por exemplo, permite identificar proteínas biomarcadoras da doença, marcadores de prognóstico e uma melhor compreensão da evolução tumoral.

A caracterização de sistemas biológicos por proteínas é predominantemente possível graças à espectrometria de massa [5, 6]. Esta permite analisar íons na fase gasosa. A sua aplicação à proteômica tornou-se possível graças a dois cientistas que compartilharam o prêmio Nobel de química em 2002; são eles, Dr. John B. Fenn e Kouichi Tanaka. Em particular o primeiro criou um método de ionização de cadeias poli-peptídicas (e.g., proteínas) conhecido como *electrospray* [7], utilizado para gerar os resultados hora apresentados.

Uma das técnicas proteômicas consideradas estado da arte é o *Multi-dimensional protein identification technology* (MudPIT); esta foi criada no laboratório do Dr. John R. Yates III [8]. Neste processo de identificação, uma amostra de proteínas é primeiramente digerida, formando componentes menores chamados peptídeos. Em

seguida, estes são separados por uma dupla cromatografia e injetados, gradativamente, em um espectrômetro de massa. Dentro do espectrômetro de massa estes peptídeos são fragmentados em um processo conhecido como espectrometria de massa em *tandem*.

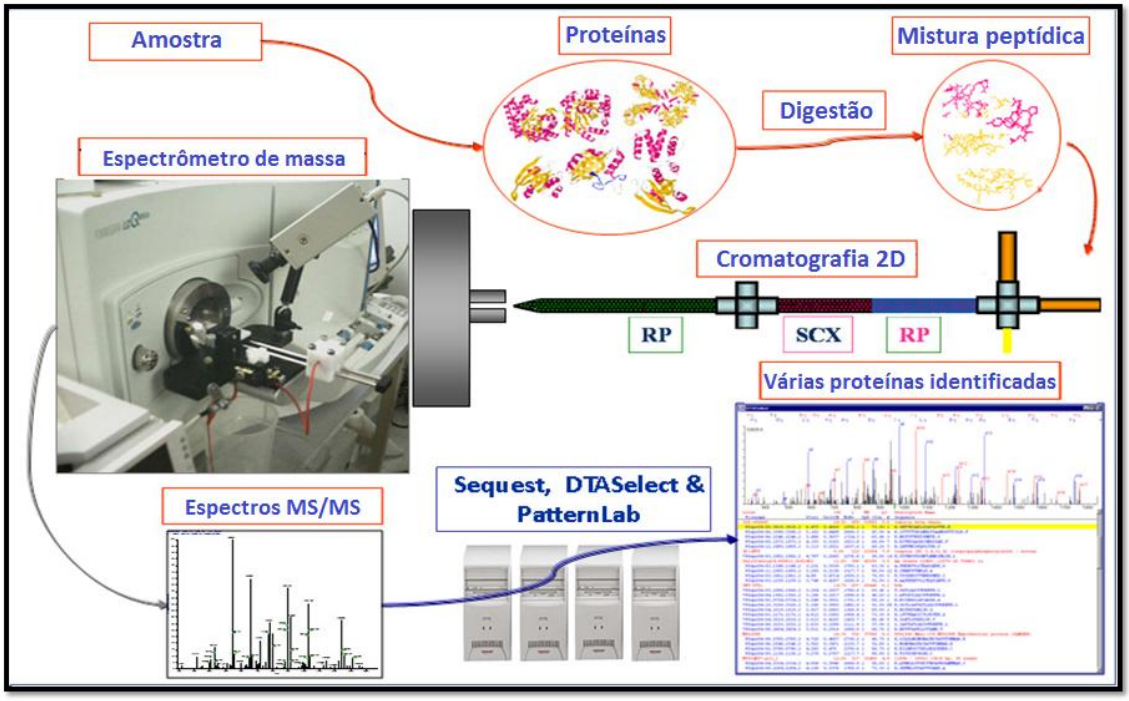

Figura 3 - Passos para identificação proteômica – Imagem cedida pelo Dr. Paulo Carvalho.

A fragmentação dos peptídeos ionizados ocorre a partir da colisão dos mesmos com um gás inerte (como moléculas de nitrogênio ou átomos de argônio ou hélio) e entre eles mesmos. A informação das massas destes fragmentos (e.g., espectro de massa em *tandem*) serve como entrada para que algoritmos realizem a identificação de qual proteína o respectivo peptídeo originou-se.

Em geral, os íons de maior intensidade são aqueles provenientes de moléculas onde ocorreu uma quebra da ligação peptídica (i.e., entre aminoácidos). Esta classe de íons é denominada de íons do tipo **b**, quando o fragmento do peptídeo íon é proveniente da parte oriunda do nitrogênio (N) terminal, e do tipo **y** quando do carbono (C) terminal do peptídeo parental conforme demonstrado na Figura 4.

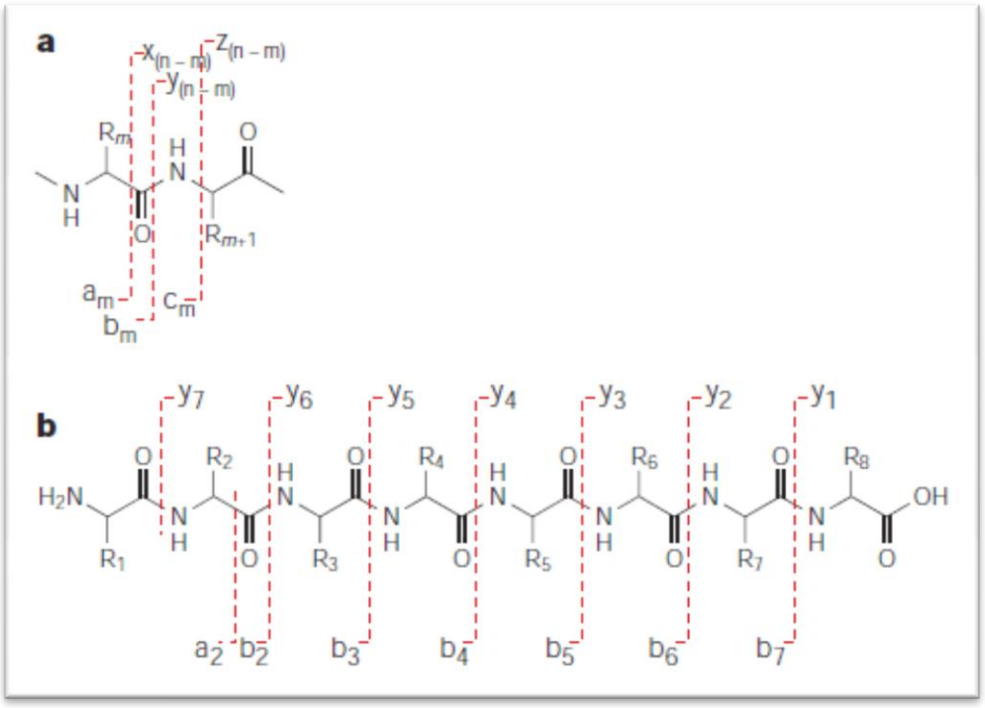

Figura 4 - A Figura 4a mostra a estrutura química de um peptídeo e onde ocorre a fragmentação do mesmo. Íons são rotulados a partir da terminação amino original a_m, b_m e c_m na qual m representa o número de grupos de aminoácidos R que estes íons contêm. Eles também são rotulados a partir da terminação carboxi original $z_{(n-m)}$, $y_{(n-m)}$ e $x_{(n-m)}$, na qual n-m representa o número de grupos R que estes íons contêm (n é o número de grupos R e m o número que a-, b- ou c-íon correspondente deveria conter. Figura obtida de [5].

Uma vez obtidos esses espectros experimentais, a tarefa é utilizar os mesmos para identificar quais proteínas estão presentes na amostra. Isso é feito através da comparação dos espectros experimentais com espectros teóricos gerados a partir de um banco de dados de sequências de proteínas; existem algoritmos eficientes que realizam esta tarefa [9, 10, 11]. A geração dos espectros teóricos envolve realizar uma digestão *in silico* das sequências protéicas do banco de dados e, em geral, para cada peptídeo, listar as massas de todos os íons das series *b* e *y*. *Scores* de similaridade entre o espectro experimental e o teórico são calculados e são utilizados como indicativos de uma confiança da identificação. Este processo é computacionalmente custoso, muitas vezes sendo executado em *clusters* de computadores.

Muitos dos espectros não geram identificações protéicas. Isto acontece por diversos motivos, como: podem estar presentes outros componentes químicos na

amostra analisada que não eram esperados pelo banco de dados, pela possibilidade da proteína ter sofrido uma modificação pós-traducional (e.g., estar fosforilada) e esta não estar prevista nos parâmetros de busca, por ter peptídeos que não foram eficientemente fragmentados e por motivos desconhecidos. Independentemente da qualidade do espectro, o mesmo servirá como entrada para a ferramenta de busca e ela proverá a sequência peptídica mais provável.

Os algoritmos atuais para proteômica usam uma estratégia chamada *target-decoy*. Nesta estratégia o banco é criado a partir de sequências de proteínas-alvo para a identificação (*target*) e de sequências artificiais de proteínas inexistentes no proteoma-alvo (*decoy*) que servem como referência interna de erro. Estas sequências artificiais preservam a frequência original dos aminoácidos do organismo. Geralmente, as sequências *decoy* são geradas invertendo as sequências do *target*, embora outros grupos a gerem de forma estocástica. Os números de sequências *target* e *decoy* costumam estarem presentes na mesma frequência [12, 13].

À medida que o espectrômetro de massa vai gerando espectros da amostra analisada, estes precisam ser comparados com espectros teóricos gerados a partir do banco de dados de sequências de proteínas a fim de identificar as proteínas presentes na amostra. Esta tarefa é feita por um programa de busca que utiliza métricas de similaridade para comparar os espectros teóricos gerados a partir do banco de dados. Cada espectro vai casar com uma lista de sequências de peptídeos (*peptide sequence match* – PSM). Este programa ordena os PSMs por um *score* de similaridade principal (e.g., *cross correlation*) e fornece o resultado de outros *scores* de confiança. Para cada espectro, um ou mais PSMs são fornecidos. Apenas o PSM de maior *score* de similaridade principal escolhido é considerado para análise posterior.

A partir dos *scores* de similaridade e do rótulo de cada espectro (i.e., *decoy* ou *target*) é gerado um modelo de classificação com duas classes. Este discriminador tem como objetivo fornecer um *score* que está correlacionado com a confiança da identificação do respectivo PSM. Os PSMs então são ordenados de forma decrescente por este último *score*.

Em seguida é definida uma pontuação de corte para o *score* de confiança de forma que a lista só seja formada por espectros com valores acima deste limite. Este limite é escolhido de tal forma que só estejam presentes apenas 1% de *decoys* no resultado final. A comunidade proteômica aceita 1% de "descobertas falsas" (*False Discovery Rate*, FDR) no resultado de um experimento [12]. Exemplos de *softwares* que seguem estes passos usando diferentes modelos de classificação são DTASelect [15] e IDPicker [14].

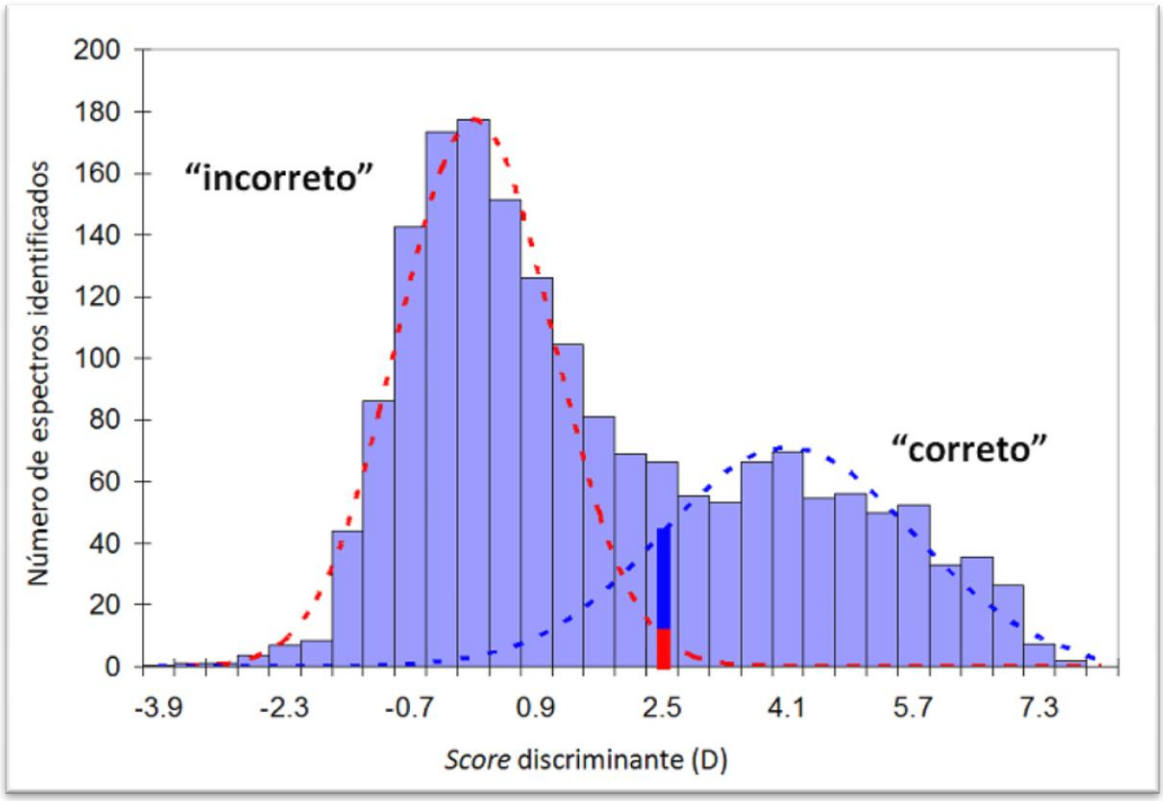

Figura 5 - Histograma de espectros identificados. A gaussiana vermelha representa identificações incorretas e a gaussiana azul as identificações corretas. Figura obtida de [16].

A Figura 5 mostra a abordagem utilizada pelo programa *Scaffold* da *Proteome Software*; resumidamente, a figura mostra um histograma dos PSMs, onde o eixo das ordenadas representa o número de PSMs e o das abscissas representa os *scores*. Este gráfico apresenta uma distribuição bimodal, onde é traçada uma gaussiana vermelha para representar as identificações incorretas e uma gaussiana azul para representar as identificações corretas.

Ao se definir um limite de 1% para o número de *decoys* no resultado, é assumido que teremos aproximadamente o mesmo número de identificações incorretas escondidas com o *target*. Isto é apropriado apenas se existir uma probabilidade igual de selecionar um casamento de peptídeo incorreto nas sequências *target* e *decoy* [12].

Neste trabalho, mostramos que nem sempre essa abordagem garante que o FDR de 1% seja uma representação fidedigna da realidade. Ou seja, o 1% de FDR pode não se traduzir em um erro real de 1%; o que pode estar ocorrendo é na verdade o produto do sobreajuste (*overfitting*) do modelo de classificação.

Para apontar estas situações onde o FDR pode induzir o pesquisador a um resultado equivocado, propomos um novo método aqui intitulado de semirrotulado, para avaliar a confiança do resultado das identificações, o qual será detalhado mais a frente.

Nossa metodologia é importante, pois devido à competição acirrada entre grupos acadêmicos, cada vez mais os programas de filtragem utilizam funções discriminadoras mais complexas objetivando maximizar as identificações com um mesmo FDR. A técnica padrão ouro não contabiliza o sobreajuste do discriminador aos dados. Para contornar esta situação, criamos um método estatístico para avaliar o sobreajuste. Nosso método foi testado em um lisado de *Pyrococcus furiosus* e um lisado de *Tripanosoma cruzi* analisados por *MudPIT*. Este trabalho aponta uma falha no método padrão ouro e

este problema nos motivou a buscar uma solução para contorná-lo e poder oferecer um resultado confiável para identificações protéicas.

1.2 – Reconhecimento de padrões

Para executar a tarefa de filtragem de espectros este trabalho utilizou técnicas de reconhecimento de padrões [17]. Resumidamente, reconhecer um padrão é saber identificar características que se repetem. Um exemplo de reconhecimento de padrões são os filtros de *spam* para *emails*: a partir de características presentes nas mensagens, como o número de destinatários e outras configurações opcionais, as mensagens são classificadas nas devidas categorias. Quanto mais dados tivermos, mais preciso será o processo de classificação.

Uma máquina também pode ser capaz de reconhecer padrões. Esta tarefa é feita a partir da construção de modelos que as capacitam a aprender e inferir a partir de experiências passadas ou exemplos. Logo, similarmente a uma pessoa, as máquinas precisam "aprender" sobre as diferentes classes em um universo de exemplos. Através deste processo de aprendizado elas poderão classificar novas entradas com maior precisão.

As três formas mais conhecidas de aprendizado de máquina são:

1. Supervisionado: os exemplos são rotulados, ou seja, já se sabe a priori as classes do problema e a que classe cada exemplo pertence.
2. Semi supervisionado: apenas uma fração dos exemplos estão rotulados. Esses exemplos rotulados trabalham em conjunto com os não rotulados para criar uma função discriminante de cada classe.
3. Não supervisionado: não existem exemplos rotulados. O treinamento consiste basicamente em formar grupos com características aproximadas.

O problema abordado neste trabalho forneceu um conjunto de exemplos rotulados. Dado isto, duas metodologias de classificação foram avaliadas para validar o novo método de filtragem de espectros. Foram eles: o classificador Bayesiano [17] e a WiSARD [18, 19].

O classificador Bayesiano é um método de classificação que assume que os dados a serem analisados possuem uma distribuição normal e modela uma função discriminante para classificar futuros exemplos. Embora os dados nem sempre possuam uma distribuição normal, este classificador tem se mostrado robusto mesmo assim [20, 21, 22].

O modelo WiSARD (***Wi****lkes,* ***S****tonham and* ***A****leksander* ***R****ecognition* ***D****evice*) é uma rede neural sem peso baseada em um rede de nós de memória de acesso aleatório (*Random Access Memory* – RAM *nodes*) [18, 19]. A rede anteriormente referida foi inspirada em como sinalizações inibitórias e excitatórias ocorrem em uma árvore dendrítica. Uma simples analogia existe entre a prioridade do sinal associada com a altura das conexões dos neurônios pré-sinápticos à árvore dendrítica e como a decodificação de endereços binários é feita em memórias de acesso aleatório (RAM).

O Capítulo 2 define os objetivos deste trabalho. O Capítulo 3 detalha como foram obtidos os dados para que cada objetivo fosse atingido. O Capítulo 4 apresenta todos os resultados obtidos em figuras e tabelas ilustrativas. O Capítulo 5 possui uma discussão dos métodos utilizados e os resultados obtidos. O Capítulo 6 apresenta as conclusões deste trabalho. Finalmente o Capítulo 7 enumera as referências bibliográficas utilizadas. Esta dissertação é finalizada pelo Anexo I que possui o resumo em inglês e a resposta do *paper* com os resultados deste trabalho submetido à PROTEOMICS.

2 – Objetivos

O objetivo principal deste trabalho é a criação de uma metodologia para estimar o sobreajuste dos classificadores aos dados proteômicos. Os objetivos específicos encontram-se a seguir:

2.1 – Objetivo 1: Implementação e Avaliação da WiSARD

Ganhar experiência com um classificador com parâmetros ajustáveis como a WiSARD é essencial para que se possa aplicá-lo eficientemente ao problema. Experimentos iniciais foram realizados para que esta tarefa pudesse ser efetuada.

2.2 – Objetivo 2: Implementação e Avaliação do Classificador Bayesiano

O classificador Bayesiano possui uma função discriminante que faz uso de operações matriciais. Foi importante então avaliar como esta característica pode influenciar no seu desempenho.

2.3 – Objetivo 3: Estimar o sobreajuste dos classificadores aos dados proteômicos

Após testes iniciais, o conhecimento acumulado sobre os classificadores foi o suficiente para poder estimar o sobreajuste dos mesmos aos dados proteômicos. A próxima seção descreve como estes testes foram feitos e como os dados proteômicos foram usados para o método semirrotulado.

3 - Metodologia

3.1 – Classificadores

O classificador Bayesiano e a WiSARD foram os modelos de classificação escolhidos para analisar o sobreajuste aos dados proteômicos. O uso destes é prática usual de acordo com a bibliografia e por experiência do autor.

3.1.1 – Classificador Bayesiano

Este item contém um extrato resumido do capítulo 2 da referência [17].

3.1.1.1 – Teorema de Bayes

O Teorema de Bayes é o cerne do classificador Bayesiano. Este teorema pode ser enunciado da seguinte forma [23]: Suponha que os eventos $w_1,w_2,\ldots,w_n$ formam uma partição do espaço amostral S; ou seja, os eventos w_i sejam mutuamente exclusivos e sua união é S. Seja *x* um evento qualquer. Então para qualquer *i* tem-se:

$$P(w_i|x) = \frac{P(w_i)p(x|w_i)}{P(w_1)p(x|w_1) + P(w_2)p(x|w_2) + \cdots + P(w_n)p(x|w_n)}$$

De uma maneira geral assume-se que para cada evento *i* existe uma probabilidade $P(w_i)$ *a priori* de sua ocorrência que reflete o conhecimento que se tem da freqüência esperada de sua ocorrência antes que este evento ocorra. É possível que se tenha a necessidade de tomar uma decisão sobre a natureza ou tipo de um evento que irá ocorrer com base apenas no conhecimento de probabilidades *a priori*. Neste caso, se tenho que classificar o evento como do tipo *i*, a decisão que se apresenta é: decida w_i se $P(w_i) > P(w_j)$, para todo *j*, tal que $i \neq j$.

Na maioria das vezes tem-se a necessidade de tomar decisões com um nível maior de informações. Observando a fórmula de Bayes, esta nos diz que se pode converter a probabilidade *a priori* $P(w_i)$ numa probabilidade *a posteriori* $P(w_i \mid x)$ que, na prática, é a probabilidade da existência de um evento w_i, dado que se tem a medida de uma característica x. Chamamos $p(x \mid w_j)$ de verossimilhança de w_j em relação a x, termo escolhido para indicar que, mantendo as demais variáveis fixas, o evento w_j para o qual $p(x \mid w_j)$ tem o maior valor indica ser a decisão verdadeira. No caso prático cada evento w_i corresponde a uma classe de um problema real de classificação.

3.1.1.2 – Teoria da Decisão Bayesiana

Os conceitos até aqui expostos sobre a teoria da decisão com base na análise Bayesiana permitem generalizá-los em três vertentes:

1. O uso de mais de uma característica;
2. A consideração de mais de duas classes;
3. A introdução de uma função mais geral do que a função probabilidade de erro que incorpora o conceito de quão onerosa é cada decisão.

A primeira vertente que é o uso de mais de uma característica requer a substituição de um escalar x representativo da característica por um vetor de características $\vec{x} = [x_1, x_2, ..., x_d]$ em um espaço Euclidiano *d-dimensional* R^d chamado de espaço de características. A segunda vertente, que é a consideração de mais de duas classes, permite que o problema restrito a duas classes seja expandido ao caso geral de várias classes. Finalmente, a terceira vertente permite a introdução de uma função de custo que estabelece quão oneroso é a tomada de cada decisão, uma vez que alguns

erros de classificação são mais onerosos do que outros, embora frequentemente se discuta o caso mais simples em que todos os erros são igualmente onerosos.

Como o custo desta decisão está associado ao erro, é comum tratar erro e risco com a mesma terminologia. O objetivo de um classificador é minimizar o risco no processo de classificação.

3.1.1.3 – Funções discriminantes

Uma maneira de representar classificadores de padrões é empregar um conjunto de funções $g_i(x)$, $i = 1 \ldots c$. Para o caso geral de riscos do classificador Bayesiano, nós podemos ter $g_i(x) = -R(\alpha_i \mid x)$ (risco associado a uma decisão α_i), uma vez que a função discriminante máxima corresponderá ao mínimo risco condicional. No caso do critério da taxa de erro mínima é possível simplificar a análise fazendo $g_i(x) = P(w_i \mid x)$, de tal modo que a função discriminante não é única e, de maneira geral, substitui-se $g_i(x)$ por $f(g_i(x))$, onde $f(.)$ é uma função monotonicamente crescente, sem que o resultado da classificação seja alterado, observação que leva a simplificações analíticas e computacionais. No caso particular da classificação por taxa de erro mínima, as seguintes funções podem ser usadas:

- $g_i(x) = P(w_i \mid x) = \frac{p(x \mid w_i)P(w_i)}{\sum_{j=1}^{c} p(x \mid w_j)P(w_j)}$;
- $g_i(x) = p(x \mid w_i)P(w_i)$;
- $g_i(x) = \ln p(x \mid w_i) + \ln P(w_i)$;

onde ln representa o logaritmo natural.

3.1.1.3.1 – O caso de duas classes

Ao invés de se usar duas funções $g_1(x)$ e $g_2(x)$ é mais comum se utilizar um único discriminante na forma: $g(x) = g_1(x) - g_2(x)$. Como regra de decisão utilizar: decida w_1 se $g(x) > 0$; w_2 caso contrário.

3.1.1.3.2 – O caso de mais de duas classes

Uma das funções discriminantes mais usuais tem a forma $g_i(x) = \ln p(x \mid w_i) + \ln P(w_i)$. A distribuição normal tem, para a teoria das probabilidades, uma importância maior do que qualquer outra distribuição pelas seguintes razões apontadas em [23]:

- Normalmente ocorre em problemas práticos;
- Fornece uma aproximação precisa para um largo número de outras distribuições de probabilidade;
- Muito embora uma variável aleatória não tenha uma distribuição normal, a média aritmética de *n* observações desta variável, tem, para *n* grande, uma distribuição aproximadamente normal.

Além dessas versatilidades da distribuição normal, esta se torna especialmente atrativa quando se utiliza a função discriminante acima, pois sua natureza é exponencial, ou seja, esta função pode ser facilmente determinada se a função densidade $p(x \mid w_i)$ é uma distribuição normal simbolizada por $p(x \mid w_i) = N(\mu_i, \sigma)$ pois a função densidade para uma variável tem a forma:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right]$$

Para mais de uma variável tem a forma:

$$p(x) = \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu)^t \Sigma^{-1}(x - \mu)\right]$$

Onde *x* é um vetor coluna com *d* dimensões; μ é um vetor de médias com *d* dimensões; $\sum$ é a matriz de covariância *d* × *d*; $|\sum|$ é o determinante da matriz de covariância; $\sum^{-1}$ é o inverso da matriz de covariância; $(x - \mu)^t$ é a matriz transposta de $(x - \mu)$. Portanto, a expressão da função discriminante, neste caso, é:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

Segundo [17], a classificação resultante da aplicação desta função discriminante vai depender da matriz de covariância de acordo com três casos típicos:

1. $\Sigma_i = \mu^2 I$;

   Este é o caso mais simples em que as características são independentes e cada uma delas tem a mesma variância σ^2.

2. $\Sigma_i = \Sigma$, para todo i;

   Neste caso as matrizes de covariância de todas as classes são iguais.

3. $\Sigma_i = arbitrária$;

   Este é o caso mais geral em que as matrizes de covariância são diferentes para cada classe. Neste trabalho todos os casos estudados obedeceram a este caso.

3.1.2 - WiSARD

O modelo WiSARD é uma rede neural sem-peso formada por discriminadores compostos de RAM *nodes* (neurônios). Cada discriminador d_i consiste de *X* RAM *nodes* que possuem *n* bits de entrada e 1 bit de saída. O vetor de entrada $\vec{x}$ tem todos os seus *X.n* bits conectados, através de um mapeamento biunívoco pseudoaleatório, aos bits de entrada dos RAM *nodes*.

O vetor de entrada $\vec{x}$ pode ser obtido a partir de um conjunto de números decimais, representando as dimensões de um problema, os quais são transformados em

um vetor de $X \times n$ bits através da codificação Gray [24]. Esta codificação é usada em substituição a transformação comum de números da base decimal para a base binária a fim de manter a distância de Hamming a menor possível. Esta técnica garante que dois números consecutivos só difiram em 1 bit apenas.

Para treinar d_i, este recebe um vetor de entrada $\vec{x}$ contendo um exemplo de uma classe alvo através da transformação descrita acima. Este vetor precisa, em seguida, ser permutado para garantir um melhor funcionamento do modelo. Neste trabalho, esta operação é feita da mesma forma como um baralho é embaralhado: o vetor de bits é dividido em dois vetores e um terceiro vetor é formado começando pelo primeiro bit do primeiro vetor, recebendo em seguida o primeiro bit do segundo vetor, o segundo do primeiro vetor, o segundo do segundo vetor e assim por diante. Este processo pode ser repetido indefinidamente usando o vetor resultante de cada iteração como entrada.

Uma vez que os bits foram permutados, este novo vetor $\vec{y}$ é então usado como entrada em d_i, marcando como visitadas todas as posições de memória endereçadas pelo vetor, isto é, um é escrito naquelas posições (todas as RAM *nodes* têm como conteúdo inicial zero). Este processo é repetido para cada vetor de entrada recebido por d_i . Outros discriminadores, cada um representando uma classe diferente, são treinados da mesma forma.

A resposta $g_i(x)$ de um discriminador é obtida através da soma de todas as saídas pertencentes ao d_i em relação a um vetor $\vec{x}$ de teste, tal que:

$$g_i(x) = \sum_{0}^{X-1} R_i^j$$

onde R_i^j representa a saída do *j-ésimo* nó RAM do discriminador d_i .

A Figura 6 ilustra como funciona o mapeamento dos bits de entrada e como é feita a decisão da classe de teste na WiSARD. Observe que o valor de confiança DS é definido como a diferença entre os dois discriminadores com o maior valor de resposta.

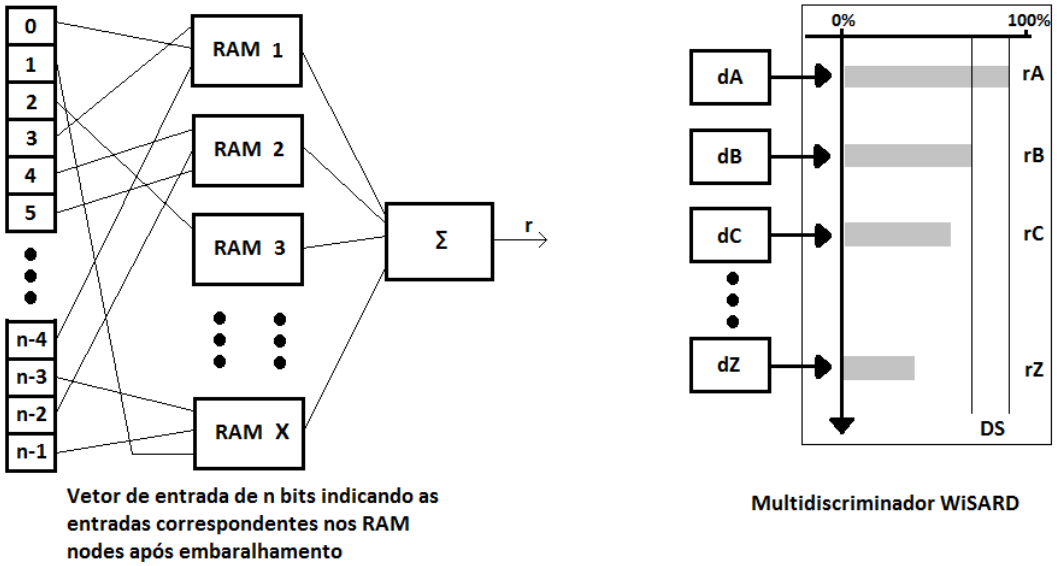

Figura 6 - Representação gráfica do modelo WiSARD

À medida que o tamanho do conjunto de treinamento aumenta, é esperado que dois ou mais discriminadores possam dar como saída o maior valor possível (saturação), isto é, $g_i(\vec{x}) = X$, quando receberem vetores de teste. Para lidar com este problema, uma simples generalização da arquitetura da WiSARD é usada [19]: cada posição de um RAM *node* tem um contador, tal que, no início do processo de treinamento, cada contador é inicializado com valor zero e é incrementado toda vez que a mesma posição é visitada. Depois que o processo de treinamento é feito os contadores possuem valores iguais ou maiores que zero.

Um valor limite mínimo *b* é aplicado em todos os RAM *nodes* antes que os discriminadores calculem $g_i(x)$. Começando por $b = 1$, todas as respostas dos discriminadores são calculadas; se DS for zero, isto é, se ocorrer um empate, *b* é incrementado e todas as respostas dos discriminadores são calculadas novamente. Este

procedimento é repetido até que DS forneça uma resposta maior que zero ou que todos os discriminadores respondam zero [18].

3.2 – Testes iniciais

O classificador Bayesiano e a WiSARD foram submetidos a testes iniciais como um estudo do comportamento dos dois métodos em relação a diferentes problemas que serão descritos a seguir. O objetivo era poder entender as características e comportamentos de cada um e adquirir experiência para aplicá-los aos dados proteômicos. Estes classificadores foram implementados usando a linguagem de programação JAVA.

O primeiro teste criado foi um aplicativo onde o usuário tem a opção de marcar um ponto com uma entre três cores a escolher. As coordenadas *X* e *Y* dos pontos são as informações que formam os vetores de entrada e as cores a classe de cada vetor. Esses dados são usados como treinamento em um dos métodos. Após o processo de treinamento, o quadro é analisado pixel por pixel, usando as coordenadas do pixel como vetor de teste. Dada a resposta dos discriminadores, o pixel é colorido com a cor correspondente. Uma ferramenta foi construída para ilustrar o comportamento dos dois classificadores com este tipo de problema.

Tendo adquirido uma experiência maior com cada classificador, decidimos lidar com um problema real que foi a aplicação dos métodos em um conjunto de dados de dígitos escritos manualmente (base MNIST <http://yann.lecun.com/exdb/mnist/>). Eles foram gerados a partir de imagens de tamanho 28×28 pixels, cada pixel com um nível de cinza entre 0 e 255.

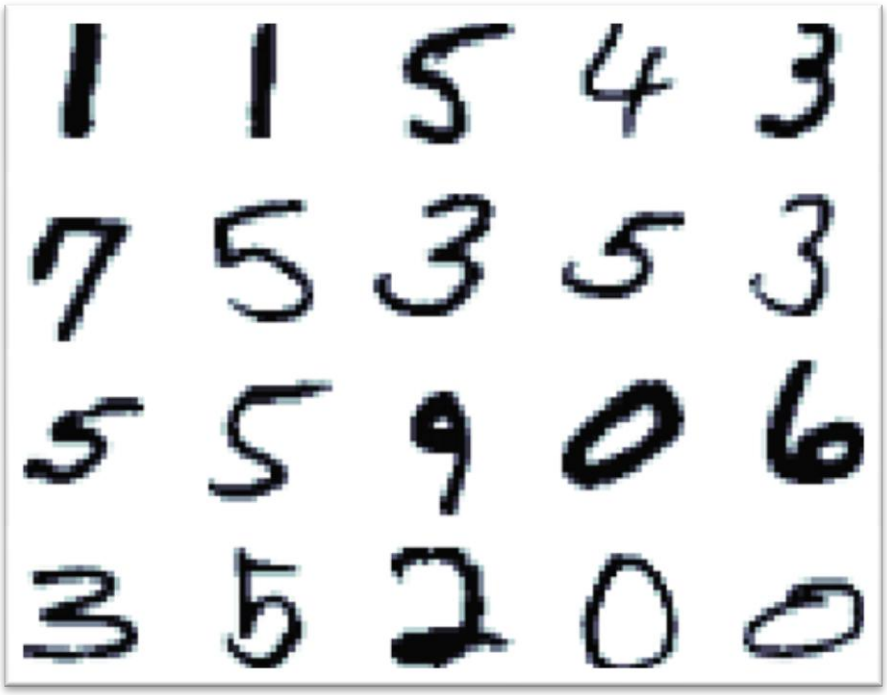

Figura 7 - Exemplos de imagens de dígitos escritos à mão. Figura obtida de [25].

Os vetores de entrada são constituídos por 784 dimensões, cada dimensão contendo o nível de cinza de cada pixel da imagem. O problema possui dez classes, que são os dígitos de zero a nove. A base de dados fornece um conjunto de treinamento com 60 mil exemplos e um conjunto de teste com 10 mil exemplos.

Esta base MNIST foi usada também para avaliar como a permutação dos bits de entrada pode influenciar no resultado de classificação. Para isso foi usado *10-fold cross validation* com o número de permutações variando de um a cem. Neste processo o conjunto de treinamento foi dividido em dez subconjuntos, dos quais nove subconjuntos eram usados para o treinamento da WiSARD e um era usado para teste. Este procedimento é feito para cada subconjunto até que todos sejam testados.

3.3 – Dados proteômicos

Os métodos avaliados finalmente puderam ser usados para avaliar o sobreajuste dos classificadores aos dados proteômicos. Primeiramente amostras de proteínas de

Pyrococcus furiosus e *Trypanosoma cruzi* foram analisadas por MudPIT; o primeiro lisado foi obtido do laboratório do Dr. John R. Yates III no *The Scripps Research Institute* em La Jolla, Califórnia – EUA, e o segundo do Laboratório de Toxinologia do Instituto Oswaldo Cruz do Rio de Janeiro. Ambas representam uma mistura altamente complexa de proteínas, o que permite avaliar as metodologias de classificação assim como o método semirrotulado.

Os dados usados foram obtidos a partir de espectros de massa gerados por espectrômetros de massa de última geração, o Orbitrap XL *(Thermo, San Jose, CA*), que serviram como entrada para o programa de busca ProLuCID [9]. Resultados onde a identificação divergia em mais de doze partes por milhão (PPM) da sequência teórica foram eliminados.

Cada resultado da ferramenta de busca foi representado como um vetor de entrada formado pelos *scores* de confiança fornecidos pela ferramenta de busca e o respectivo rótulo. Estes *scores* de confiança são: *cross correlation*, *binomial score*, *DeltaCN***,** número de picos (i.e., razão massa carga de íons apontados no espectro) casados dividido pelo número de picos total, PPM, número de terminações trípticas, *ranking* secundário e *z-score*. Estes vetores são usados para construir um modelo de classificação que vai fornecer um novo *score* de confiança para cada espectro.

Define-se o número de terminações trípticas de um peptídeo quando ambas as extremidades do peptídeo foram clivadas pela enzima em questão. A tripsina é uma enzima que cliva cadeias polipeptídicas após arginina e lisina. O número de terminações trípticas (*fully tryptic, half tryptic, non-tryptic*) é o número de clivagens após arginina e lisina, não podendo ser antes de prolina.

O capítulo anterior apresentou o método padrão ouro na proteômica e este trabalho tem como objetivo mostrar que existe uma falha neste método. Apresentamos

uma nova abordagem no método de identificação de proteínas. Esta abordagem, chamada método semirrotulado, mantém no banco de dados de sequências de proteínas as sequências *target* e apresenta uma nova maneira de analisar os resultados.

Os dados obtidos pela ferramenta de busca são usados como vetores de entrada para a construção de um modelo de classificação. O banco de dados no nosso novo método contém sequências *decoy* e, adicionalmente, um conjunto de sequências *decoy* que o classificador não sabe que são *decoys* (*unlabeled decoys*).

Criado o modelo, este é usado para calcular a confiança de cada PSM e em seguida estes PSMs são ordenados por este *score* de confiança, assim como é feito no método *decoy*. Ao aplicar um FDR de 1%, é possível analisar a proporção de *decoys* e *unlabeled decoys*. Para que o resultado seja de confiança espera-se que a quantidade dos dois seja aproximadamente a mesma.

Em uma primeira tentativa (*target, reverse e scrambled,* T-R-S) para a construção das sequências *decoy* e *unlabeled decoy*, a estratégia de sequência reversa foi preservada para o *decoy*. Para o *unlabeled decoy* foi escolhida a opção de simplesmente embaralhar (*scrambled*) as sequências *target*. Porém esta opção não garante a validade do método já que as sequências embaralhadas não preservam a redundância na base de dados. Existem peptídeos que são comuns a várias proteínas e quando um sistema de embaralhamento para geração das sequências de proteínas é utilizado, a diversidade dos peptídeos presentes no banco de dados aumenta em relação às sequências *target*.

Uma segunda opção (*target, scrambled0, scrambled1,* T-S0-S1) foi, além de embaralhar os *unlabeled decoy*, embaralhar também os *decoys*. Apesar de alguns grupos aceitarem o método de embaralhamento de sequências para gerar *decoys*, conforme descrito acima, o número de peptídeos *decoy* ficará muito superior ao de peptídeos

target; isto vai gerar uma comparação desbalanceada durante a criação do discriminador.

Finalmente foi criada uma opção (*target, pair reversed e middle reversed*, T-PR-MR) para reproduzir as vantagens presentes na estratégia de inverter as sequências *target*. Nela duas sequências de proteína são geradas para cada *target*: *Pair Reversed* (PR) e *Middle Reversed* (MR); antes de estas sequências serem geradas é necessário informar a enzima de digestão usada no projeto. Para cada sequência *target*, o algoritmo primeiro gera a lista de peptídeos que a enzima selecionada produziria. A sequência PR é gerada pegando cada peptídeo do *target* e esta sequência de peptídeo é invertida e os aminoácidos são trocados de posição dois a dois, exceto nas terminações: a sequência ABCDEFGHI torna-se IGHEFCDBA. A sequência PR final é obtida concatenando todos os peptídeos PR. A sequência MR é gerada pegando cada peptídeo *target* e trocando os aminoácidos das pontas. A parte interna é obtida dividindo o aminoácido ao meio e invertendo suas ordens. A sequência ABCDEFGHI torna-se IEDCBHGFA. Da mesma forma a sequência MR final é obtida concatenando todos os peptídeos MR.

Embaralhando-se aleatoriamente cada sequência de forma independente gera-se uma maior diversidade de peptídeos do que invertendo cada sequência, especialmente em proteomas de organismos mais complexos, com número elevado de domínios conservados, como os mamíferos. Dessa forma, a ferramenta de busca vai comparar cada espectro com mais candidatos das sequências aleatórias do que das sequências *target*. Isto vai gerar um resultado tendencioso ao estimar o FDR já que a maioria das ferramentas de busca não vai considerar o número de peptídeos distintos gerados a partir de cada base de dados de sequências de proteínas e a maioria dos cálculos de FDR assume o número de comparações de *target* e *decoys* igual. O formato T-PR-MR foi criado testando empiricamente maneiras diferentes de rearranjar as sequências de forma

a minimizar a sobreposição de picos gerados pelos espectros teóricos correspondentes ao *target*, *pair reversed* e *middle reversed*. Com essas ressalvas o método semirrotulado finalmente pôde ser validado.

O método semirrotulado consegue mostrar que nem sempre a quantidade de *decoys* e *unlabeled decoys* obedece ao esperado, e assim, aponta quando um algoritmo de filtragem apresenta um resultado equivocado.

4 – Resultados

4.1 – Resultados dos Testes Iniciais

A WiSARD se mostrou muito sensível a escolha dos parâmetros, isto é, número de bits nos RAM *nodes* e o embaralhamento do vetor de entrada. Diminuindo o número de entradas nos RAM *nodes*, é possível observar uma maior capacidade de generalização, mas à medida que o número de entradas aumenta, consequentemente diminuindo o número de RAM *nodes*, o modelo tende a gerar áreas de indecisão cada vez maiores e informações mais concentradas já que o nível de generalização está diminuindo, contribuindo para o sobreajuste.

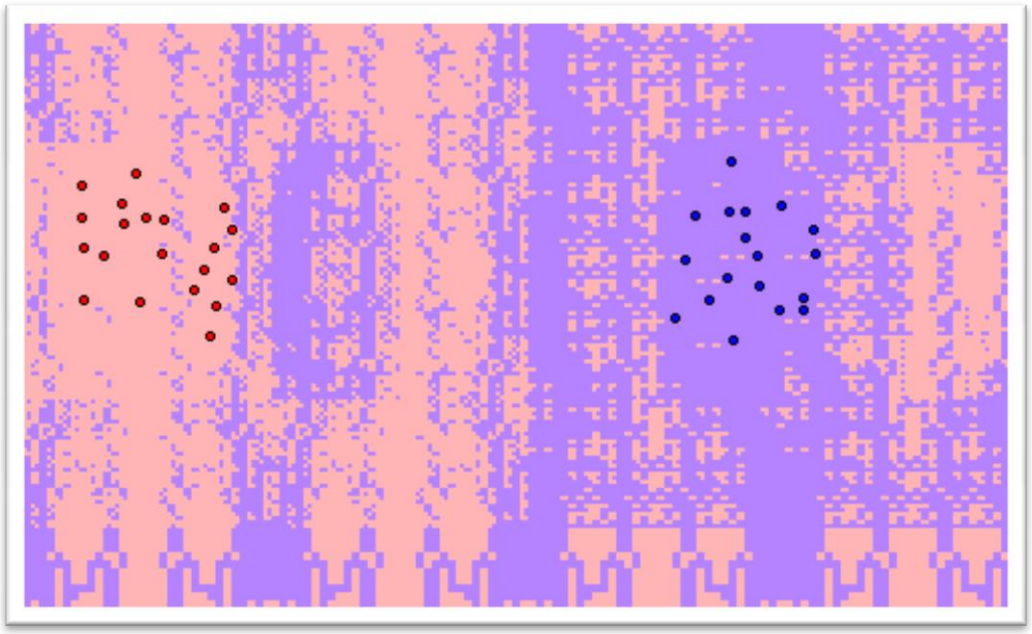

Figura 8 – Superfície de decisão gerada com a WiSARD com 8 bits de entrada nos RAM *nodes* e vetor de entrada embaralhado 7 vezes. Os pontos da figura representam o conjunto de treinamento. Os pontos vermelhos representam elementos da classe vermelha e os azuis os elementos da classe azul. A superfície rosa foi classificada como pertencente à classe vermelha e a azul claro como pertencente à classe azul.

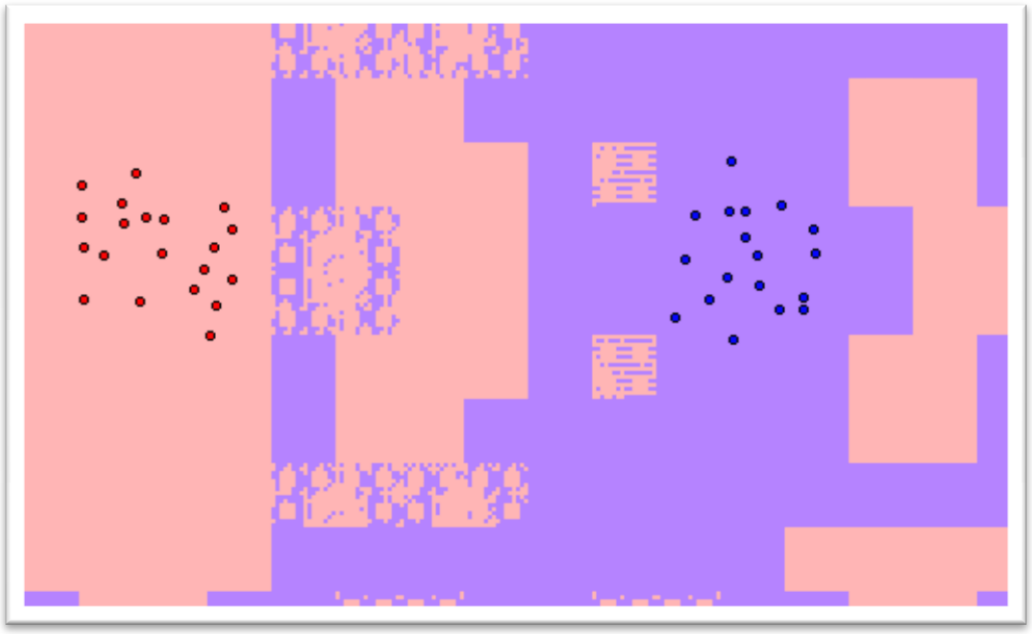

Figura 9 - Superfície de decisão gerada com a WiSARD com 4 bits de entrada nos RAM nodes e vetor de entrada embaralhado 7 vezes. Os pontos da figura representam o conjunto de treinamento. Os pontos vermelhos representam elementos da classe vermelha e os azuis os elementos da classe azul. A superfície rosa foi classificada como pertencente à classe vermelha e a azul claro como pertencente à classe azul.

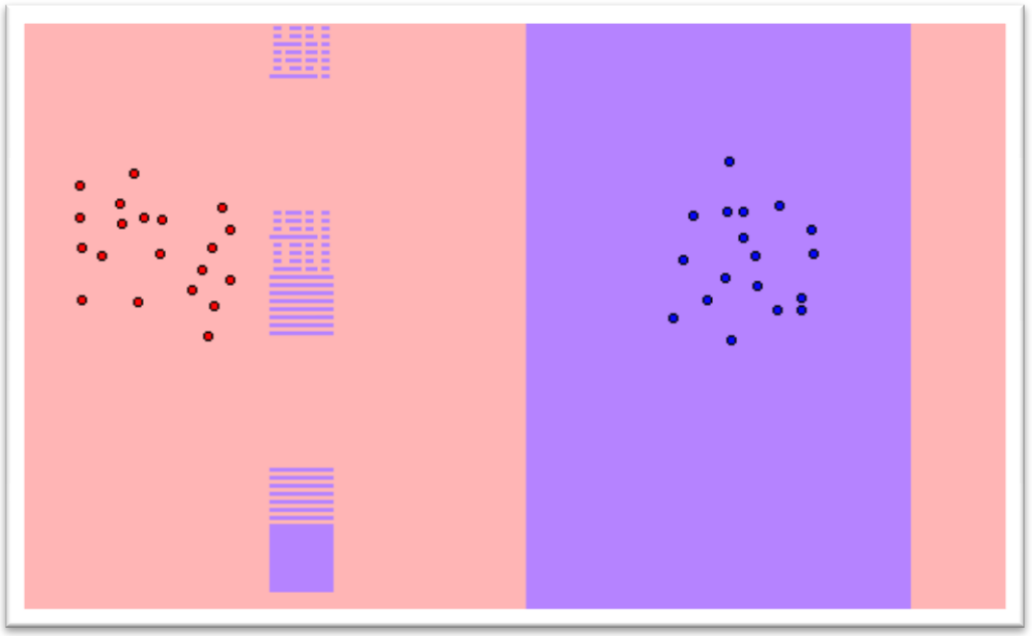

Figura 10 - Superfície de decisão gerada com a WiSARD com 2 bits de entrada nos RAM nodes e vetor de entrada embaralhado 7 vezes. Os pontos da figura representam o conjunto de treinamento. Os pontos vermelhos representam elementos da classe vermelha e os azuis os elementos da classe azul. A superfície rosa foi classificada como pertencente à classe vermelha e a azul claro como pertencente à classe azul.

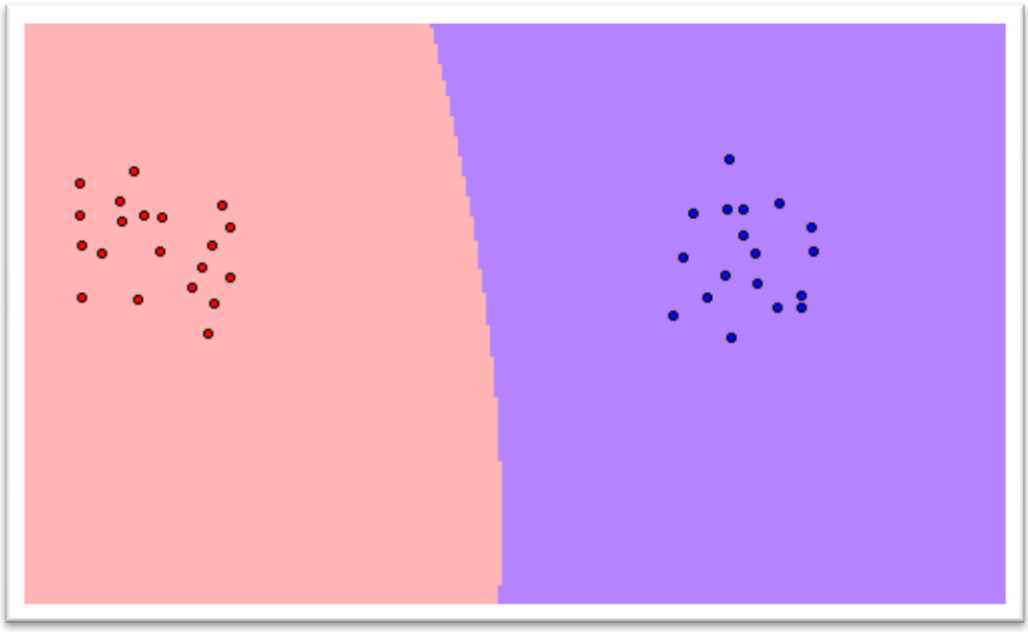

Figura 11 – Superfície de decisão gerada pelo classificador Bayesiano. Os pontos da figura representam o conjunto de treinamento. Os pontos vermelhos representam elementos da classe vermelha e os azuis os elementos da classe azul. A superfície rosa foi classificada como pertencente à classe vermelha e a azul claro como pertencente à classe azul.

No método Bayesiano, quando alguns pontos são marcados longe das médias das coordenadas X e Y da mesma classe, o classificador se mostra sensível a essa situação. Para isto foi estudado um método de eliminação dos mesmos. Este método calcula a distância de Mahalanobis ($MD_{\vec{x}}$) de um vetor de entrada $\vec{x}$

$$MD_{\vec{x}} = ((\vec{x} - \vec{m})^T \times \sum^{-1} \times (\vec{x} - \vec{m}))^{1/2} ,$$

onde $\vec{m}$ é o vetor de medianas e $\sum^{-1}$ é o vetor de covariâncias. O quadrado da distância de Mahalanobis é usado para calcular a função de distribuição cumulativa (fdc) da distribuição chi-quadrado com dois graus de liberdade. Se $\vec{x}$ tiver um fdc maior que 0,975 ele será considerado um *outlier* e removido do novo cálculo das médias e matriz de covariâncias. A Figura 12 e a Figura 13 ilustram o funcionamento do método.

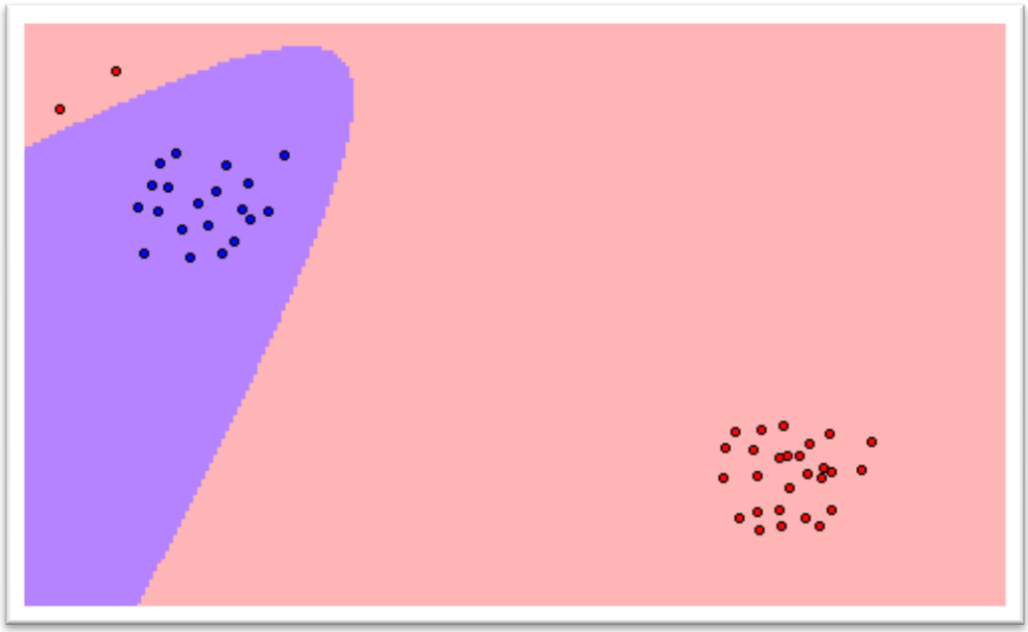

Figura 12 - Superfície de decisão gerada pelo classificador Bayesiano sem detecção de outliers. Os pontos da figura representam o conjunto de treinamento. Os pontos vermelhos representam elementos da classe vermelha e os azuis os elementos da classe azul. A superfície rosa foi classificada como pertencente à classe vermelha e a azul claro como pertencente à classe azul.

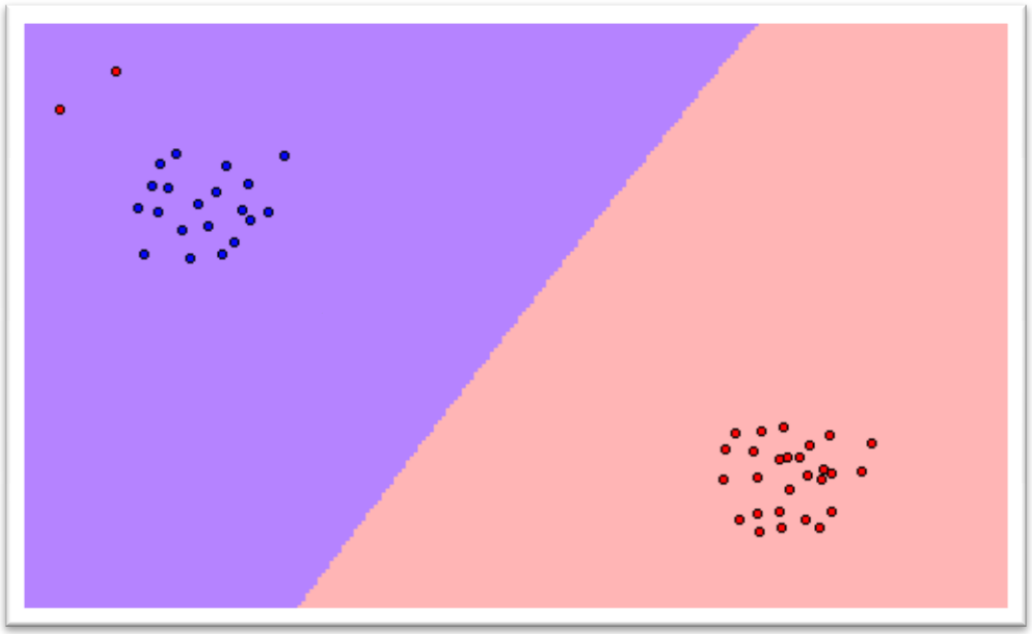

Figura 13 - Superfície de decisão gerada pelo classificador Bayesiano sem detecção de outliers. Os pontos da figura representam o conjunto de treinamento. Os pontos vermelhos representam elementos da classe vermelha e os azuis os elementos da classe azul. A superfície rosa foi classificada como pertencente à classe vermelha e a azul claro como pertencente à classe azul.

No experimento da base MNIST com a WiSARD, o treinamento do modelo foi rápido e nenhuma adaptação foi necessária. Já o classificador Bayesiano teve algumas

dificuldades de trabalhar com uma base de dados como esta. Como este problema possui 784 dimensões, as operações de inversão e multiplicação matricial tornavam-se muito lentas. Este problema precisou ser contornado para se chegar a um resultado com o classificador: as dimensões com média menor do que 30 foram eliminadas. A tabela a seguir mostra a frequência de acertos da base de teste para cada um dos métodos. O embaralhamento de 73 vezes foi o valor que obteve o melhor resultado no *cross validation* da WiSARD.

Tabela 1 – Resultados com a base MNIST. A primeira coluna representa as classes do problema; a segunda o resultado do 10 fold cross validation da WiSARD; a terceira o resultado da WiSARD com embaralhamento de 73 vezes e com 28 bits de entrada nos RAM nodes; a quarta os resultados obtidos com o classificador Bayesiano.*

Classe	WiSARD (*cross validation)*	WiSARD (embaralhamento 73, 28 bits)	Bayesiano
0	96,72%	97,86%	88,47%
1	80,52%	80,44%	63,88%
2	93,84%	93,12%	89,92%
3	89,45%	91,98%	88,02%
4	88,01%	89,30%	90,02%
5	86,56%	87,67%	78,03%
6	94,91%	94,26%	90,50%
7	82,01%	81,03%	91,15%
8	90,54%	91,68%	95,48%
9	89,78%	88,80%	83,94%

**Base de dados de dígitos escritos à mão.*

Como no problema anterior, a WiSARD se mostrou sensível ao número de permutações e esta base de dados foi usada para testar exaustivamente este parâmetro de modo a mostrar o comportamento do modelo.

4.2 – Dados Proteômicos

As tabelas de 2 a 15 ilustram os resultados obtidos com os dois classificadores, para a base de dados de *P. furiosus*. A abreviação T-S1-S0 significa que no banco de dados de sequências de proteínas há sequências reais (T) e sequências *scrambled* (S0 e S1). A abreviação T-PR-MR representa uma base de dados com sequências *Middle Reversed* (MR) e *Pair Reversed* (PR). Finalmente a abreviação T-R representa o método padrão ouro, onde temos *target* (T) e *reversed* (R) apenas. As amostras de *P. furiosus* e *T. cruzi* geraram 262515 e 8931 espectros de massa respectivamente.

Tabela 2 – Resultado do modelo Bayesiano com P. furiosus para base T-S1-S0. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-S1-S0			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espetros	1083	1105	106184	108372
Peptídeos	982	983	19609	21574

**Base de dados formada por target, scrambled1 e scrambled0.*

Tabela 3 – Resultado do modelo Bayesiano com P. furiosus para base T-PR-MR. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-PR-MR			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	1083	1064	106229	108376
Peptídeos	936	939	19587	21462

**Base de dados formada por target, pair reversed e middle reversed.*

Tabela 4 – Resultado do modelo Bayesiano com P. furiosus para base T-R. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-R		
	Decoys	*Target*	Total
Espectros	1073	106280	107353
Peptídeos	900	19532	20432

**Base de dados formada por target e reverse.*

Tabela 5 – Resultado do modelo Bayesiano com T. cruzi para base T-S1-S0. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-S1-S0			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	12	43	1221	1276
Peptídeos	9	12	267	288

**Base de dados formada por target, scrambled1 e scrambled0.*

Tabela 6 – Resultado do modelo Bayesiano com T. cruzi para base T-PR-MR. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-PR-MR			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	12	20	1235	1267
Peptídeos	11	17	273	301

**Base de dados formada por target, pair reversed e middle reversed.*

Tabela 7 – Resultado da WiSARD de 16 bits nos RAM nodes e embaralhamento de 73 com P. furiosus para base T-S1-S0. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-S1-S0			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	1150	4917	108945	115012
Peptídeos	1126	4513	22714	28353

**Base de dados formada por target, scrambled1 e scrambled0.*

Tabela 8 – Resultado da WiSARD de 16 bits nos RAM nodes e embaralhamento de 73 com P. furiosus para base T-PR-MR. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-PR-MR			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	1152	4656	109440	115248
Peptídeos	1100	4291	22803	28194

**Base de dados formada por target, pair reversed e middle reversed.*

Tabela 9 – Resultado da WiSARD de 16 bits nos RAM nodes e embaralhamento de 73 com P. furiosus para base T-R. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-R		
	Decoys	*Target*	Total
Espectros	1162	115074	116236
Peptídeos	1099	26142	27241

**Base de dados formada por target e reverse.*

Tabela 10 – Resultado da WiSARD de 16 bits nos RAM nodes e embaralhamento de 73 com T. cruzi para base T-S1-S0. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-S1-S0			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	19	381	1567	1967
Peptídeos	13	309	514	836

**Base de dados formada por target, scrambled1 e scrambled0.*

Tabela 11 – Resultado da WiSARD de 16 bits nos RAM nodes e embaralhamento de 73 com T. cruzi para base T-PR-MR. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-PR-MR			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	21	414	1722	2157
Peptídeos	20	290	601	911

**Base de dados formada por target, pair reversed e middle reversed.*

Está evidente pelas tabelas acima que o método *target-decoy* largamente usado na proteômica favorece a WiSARD de 16 bits nos RAM *nodes* contra o modelo Bayesiano, já que o primeiro identifica mais espectros do que o segundo. Porém, introduzindo os *unlabeled decoys*, mostra-se que a quantidade de falsos positivos nem sempre aparece na mesma proporção dos *decoys* como era esperado.

A partir da criação da base de dados, espera-se que o número de *unlabeled decoys* seja aproximadamente o mesmo de *decoys* nos casos do T-S1-S0 e T-PR-MR como podemos ver nas tabelas 2 e 3, mas não é o que acontece nas tabelas 7 e 8. Dado isto, um *p*-valor para o sobreajuste pode ser aproximado por $P = \Pr(X > s) \approx$

$\sum_{t=s+1}^{n} Bin(t, n, p)$, onde *X* é uma variável aleatória indicando o número de *unlabeled decoys* identificados, *s* é o valor esperado de *X*, *n* é o total de identificações, *p* é a fração de *unlabeled decoys* esperada (para um FDR de 1%, p = 0,01) e *Bin* é a função de distribuição binomial. Pelas tabelas 2 e 3 apenas a função discriminante Bayesiana pode ser considerada confiável para a base de dados de *P. furiosus* ($P >>$ 0,05).

A distribuição binomial nos permite modelar a chance de observarmos um determinado número de eventos caso soubermos, a priori, a probabilidade destes eventos e o número de tentativas. O *p*-valor para o sobreajuste calcula a probabilidade de se encontrar mais *unlabeled decoys* do que foi observado. Se $P >>$ 0,05, o classificador não sofreu sobreajuste aos dados em questão.

O que está por trás do aparente sucesso da WiSARD, ilustrado na Tabela 3, está de certa forma relacionado com o número de bits (16) nos seus RAM *nodes*, que deram ao modelo resultante a complexidade necessária para que ocorresse o sobreajuste dos dados, isto é, uma melhor separação dos *decoys* do resto. Mesmo que a WiSARD tenha fornecido uma lista maior de espectros identificados obedecendo ao FDR de 1%, o número elevado de *unlabeled decoys* demonstra que os resultados não são tão confiáveis como no caso Bayesiano.

O número de bits igual a 16 foi escolhido para comparação com o discriminador Bayesiano, pois este valor foi o que obteve um melhor resultado em relação ao número de espectros identificados e o número de *unlabeled decoys*. Valores menores que 16 forneceram um número baixo de identificação de espectros representando um resultado não muito significativo, enquanto que valores maiores indicavam mais espectros identificados, porém o número de *unlabeled decoys* cresceu consideravelmente. As tabelas abaixo ilustram os resultados com RAM *nodes* com número de bits diferentes e embaralhamento de 73.

Tabela 12 – Resultado da WiSARD de 14 bits nos RAM nodes e embaralhamento de 73 com P. furiosus para base T-PR-MR. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-PR-MR			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	996	1429	97252	99677
Peptídeos	934	1339	17845	20118

**Base de dados formada por target, pair reversed e middle reversed.*

Tabela 13 – Resultado da WiSARD de 15 bits nos RAM nodes e embaralhamento de 73 com P. furiosus para base T-PR-MR. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-PR-MR			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	1077	2928	103717	107722
Peptídeos	1015	2703	20578	24296

**Base de dados formada por target, pair reversed e middle reversed.*

Tabela 14 – Resultado da WiSARD de 17 bits nos RAM nodes e embaralhamento de 73 com P. furiosus para base T-PR-MR. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-PR-MR			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	1242	8009	115020	124271
Peptídeos	1172	7306	26305	34783

**Base de dados formada por target, pair reversed e middle reversed.*

Tabela 15 – Resultado da WiSARD de 18 bits nos RAM nodes e embaralhamento de 73 com P. furiosus para base T-PR-MR. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-PR-MR			
	Decoys	*Unlabeled decoys*	*Target*	Total
Espectros	1265	8973	116318	126556
Peptídeos	1198	8155	27253	36606

**Base de dados formada por target, pair reversed e middle reversed.*

As Tabelas 5 e 6 mostram que a abordagem computacional do experimento em questão também pode levar ao sobreajuste ($P < 0,05$), até mesmo para discriminadores,

como o Bayesiano, que se comportam bem em outras circunstâncias. Apesar de bases de dados menores serem preferíveis, já que elas representam um custo computacional menor, os números reduzidos de espectros e de identificação de *decoys* podem comprometer a robustez estatística.

Uma maneira para aliviar o sobreajuste verificado nos casos da Tabela 5 e da Tabela 6 é a aplicação de um filtro onde são eliminadas proteínas que estão presentes em apenas um espectro. Os resultados são mostrados na Tabela 16.

Tabela 16 - Resultado do modelo Bayesiano com T. cruzi com filtro aplicado para base T-PR-MR. Cada coluna representa o número de identificações em cada tipo de sequência e coluna Total o total de identificações.*

	T-PR-MR			
	Labeled Decoys	Unlabeled decoys	Target	Total
Espectros	2	6	1211	1219
Peptídeos	1	3	249	253

**Base de dados formada por target, pair reversed e middle reversed.*

.

Os resultados da Tabela 16 agora mostram um número aceitável de *unlabeled decoys* para o FDR de 1% ($P >> 0,05$). Desta forma, aplicando filtros ainda mais rigorosos, tais como considerar apenas proteínas identificadas por pelo menos dois peptídeos, torna-se desnecessário.

Tais filtros *ad hoc* são comumente usados e se mostraram bastante efetivos. Uma variação é calibrar o FDR para que seja menos rigoroso (maior que 1%) que o desejável e em seguida filtrar os resultados. Em alguns casos isto pode acarretar em mais identificações. Porém, um classificador eficiente ainda permanece como o núcleo do algoritmo de filtro de espectros, já que ele é o responsável por sortear os resultados de acordo com a confiança. Portanto, maximizar o número de identificações depende de duas restrições: o FDR e o *p*-valor de sobreajuste.

5 – Discussão

Neste trabalho avaliamos os prós e contras de cada classificador e como eles se comportaram quando foram aplicados à análise do sobreajuste aos dados proteômicos. Os testes iniciais executados antes do objetivo real evidenciaram características que podem não estar claras para quem vai estudar os dois métodos.

A WiSARD é um classificador que possui uma facilidade de ser implementado. Quando ela se deparou com um problema com muitos exemplos de treino e um alto número de dimensões, ela mostrou ser uma ótima opção, já que o seu treinamento foi rápido e o resultado bastante satisfatório. Porém ele apresentou problemas quando os números de dimensões e de exemplos eram baixos, já que ela sofreu com o sobreajuste. Isto está diretamente ligado a natureza do classificador, ou seja, na forma como a informação é guardada nos RAM *nodes*. O ajuste dos parâmetros (número de bits nas RAM *nodes*, número de RAM *nodes* e o embaralhamento do vetor de entrada) leva a um melhor resultado, embora não haja uma regra que indique as melhores opções, tornando-o pouco intuitivo.

O classificador Bayesiano mostrou-se bastante robusto em todos os testes em que foi submetido, inclusive com os dados proteômicos. Também é de fácil implementação e a sua função discriminante possui um significado bem claro, já que é ela construída a partir do pressuposto de que os dados de treinamento possuem uma distribuição normal. Este método encontrou problemas apenas no caso da base MNIST, onde o número de dimensões era alto. Por haver operações matriciais na função discriminante, um número alto de dimensões pode ser muito custoso para o método. É preciso adaptar o algoritmo para estar preparado para lidar com matrizes singulares, já que o determinante igual a zero não permite a inversão da matriz de covariâncias.

Para os dados proteômicos usados neste trabalho, ficou claro que o classificador Bayesiano obteve resultados muito melhores. A WiSARD sofreu sobreajuste em todos os testes, ao contrário do Bayesiano, embora este também tenha tido o mesmo problema para o caso da base de dados do *T. cruzi*.

Os resultados mostrados serviram para comprovar que o novo método de identificação de proteínas proposto, o método semirrotulado, é capaz de validar a confiabilidade de um resultado. Isto não era possível com o método *target-decoy* uma vez que a ausência de *unlabeled decoys* impedia que a tarefa fosse feita por completo. O método semirrotulado é um método simples, de fácil implementação e que pode ser rapidamente incorporado pela comunidade proteômica. Uma desvantagem deste novo método é que o tempo da ferramenta de busca fica maior devido ao aumento da base de dados de sequências de proteínas.

Embora este trabalho esteja voltado para um objetivo claro, que é analisar o sobreajuste dos classificadores usados aos dados proteômicos para identificações protéicas, o método criado não precisa se prender a apenas este tópico. Qualquer situação que precise de um modelo de classificação e uma análise de sobreajuste poderia se beneficiar de uma abordagem parecida.

6 – Conclusão

6.1 – Objetivo 1: Implementação e Avaliação da WiSARD

O teste de separação dos pontos marcados na superfície foi importante para entender como o número de bits nos RAM *nodes* da WiSARD está relacionado com o poder de generalização do classificador (quanto menos bits, mais genérico). O teste com a base MNIST foi importante para testar exaustivamente o embaralhamento do vetor de entrada do discriminador e como isto influencia no resultado.

6.2 – Objetivo 2: Implementação e Avaliação do classificador Bayesiano

O teste de separação dos pontos marcados na superfície serviu para ilustrar como o classificador Bayesiano constrói uma superfície de decisão. Além disto, foi sugerido um método de detecção de *outliers*. O teste com a base MNIST foi importante para apontar que o cálculo da função discriminante do classificador Bayesiano pode ser muito custoso em um problema de alta dimensão, devido às operações matriciais presentes na função. É preciso estar atento para o problema de matrizes singulares, ou seja, com determinante igual a zero, o que exige uma estratégia para contorná-lo.

6.3 – Objetivo 3: Análise do sobreajuste dos classificadores aos dados proteômicos

Finalmente, para os dados proteômicos, o objetivo era mostrar que existe uma falha no método padrão ouro. Estas falhas foram ilustradas nas tabelas do Capítulo 4. Um método de validação da confiabilidade dos resultados foi proposto e uma função indicando um *p*-valor aproximado para o sobreajuste foi construída, a fim de detectar se este fenômeno ocorreu.

7 – Referência Bibliográfica

[1] WASINGER, V.C., CORDWELL, S.J., CERPA-POLJAK, A., et al., "Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium.", **Electrophoresis**, 16, 1990-1994, 1995.

[2] <http://bp3.blogger.com/_x-XGBwYkHpA/SIiOI6-uNUI/AAAAAAAAAJo/3_uePuHnc_Q/s1600-h/Lagarta+Papilio.jpg>.

[3] <http://1.bp.blogspot.com/_tZgpHvI7XG4/S7KDvnxufYI/AAAAAAAAASg/lAy5hMbZz8M/s1600/borboleta+azul.jpg>.

[4] <http://dceg.cancer.gov/newsletter/jul03/brainMRI.jpg>.

[5] STEEN, H., MANN, M., "The ABC's (and XYZ's) of Peptide Sequencing", **Nature Reviews | Molecular Cell Biology**, Volume 5, Setembro de 2004.

[6] COLINGE, J., BENNETT, K. L., "Introduction to Computational Proteomics" **PLoS Computational Biology,** 2007.

[7] FENN, J. B.; MANN, M.; MENG, C. K.; et al, "Electrospray ionization for mass spectrometry of large biomolecules". **Science**, 246, pp 64–71, 1989.

[8] WASHBURN, M.P., WOLTERS, D., YATES, J.R. III, "Large-scale analysis of the yeast proteome by multidimensional protein identification technology" **Nat. Biotechnol**. 19, pp 242-247, 2001.

[9] XU, T., VENABLE, J.D., PARK, S.K., et al, "ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program", **Mol. Cell. Proteomics,** 5, 2006.

[10] PERKINS, D.N., PAPPIN, D.J., CREASY, D.M., et al "Probability-based protein identification by searching sequence databases using mass spectrometry data.", **Electrophoresis,** 20, pp 3551-3567, 1999.

[11] TABB, D.L., FERNANDO, C.G., CHAMBERS, M.C., "MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis.", **J Proteome Res,** 6, pp 654-661, 2007.

[12] ELIAS, J.E. & Gygi,S.P., “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry”, **Nat. Methods 4**, pp 207-214, 2007.

[13] PENG, J., ELIAS, J.E., THOREEN, et al, “Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein identifications by mass spectrometry.”, **Proteome. Res.**, 2, pp 43-50, 2003.

[14] KALL, L., CANTERBURY, J.D., WESTON, J., et al, “Semi-supervised learning for peptide identification for shotgun proteomics datasets.”, **Nat. Methods,** 4, pp 923-925, 2007.

[15] COCIORVA, D., TABB, L., YATES, J.R., “Validation of tandem mass spectrometry database search results using DTASelect”, **Curr. Protoc. Bioinformatcis.** Chapter 13, Unit 13.4, 2007.

[16] <http://www.proteomesoftware.com/>.

[17] DUDA, R. O., HART, P. E., STORK, D. G., “Pattern Classification”, 2ª Edição, 2000.

[18] GRIECO, B.P.A., LIMA, P.M.V., GREGORIO, M., et al, “Producing pattern examples from mental images”, **Neurocomputing**, 73:1057-1064, 2010.

[19] ALEKSANDER, I., MORTON, H., **An Introduction to Neural Computing**, 1990.

[20] CARVALHO, P.C., COCIORVA, D., WONG C.C., et al, “Charge Prediction Machine: Tool for Inferring Precursor Charge States of Electron Transfer Dissociation Tandem Mass Spectra”, **Anal Chem, 2009.**

[21] CHUN-YEN, C., KUO-YOUNG, C., HONG-YUAN, M.L., “Fairing of Polygon Meshes Via Bayesian Discriminant Analysis”, **WSCG**, pp 175-182, 2004.

[22] NAKAJIMA, Y., YANG LU, SUGANO, M., et al, "A fast audio classification from MPEG coded data," **Acoustics, Speech, and Signal Processing, IEEE International Conference on Proceedings**, Vol 6, 1999 IEEE International Conference on, pp. 3005-3008, 1999.

[23] LARSON, H.J., “Introduction to Probability Theory and Statistical Inference”, 2ª Edição, 1974.

[24] GRAY, F., “Pulse Code Communication”; US patent #2,632,058. March 17th, 1953.

[25] <http://3.bp.blogspot.com/_UpN7DfJA0j4/TJtUBWPk0SI/AAAAAAAAABY/oWPMtmqJn3k/s1600/mnist_originals.png>.

Anexo I – *Paper* submetido à

PROTEOMICS

Abstract

The decoy-database approach is currently the gold standard for assessing the confidence of identifications in shotgun proteomic experiments. Here we demonstrate that what might appear to be a good result under the decoy-database approach for a given false-discovery rate could be, in fact, the product of overfitting. This problem has been overlooked until now and could lead to obtaining boosted identification numbers whose reliability does not correspond to the expected false-discovery rate. To remedy this, we are introducing a modified version of the method, termed a semi-labeled decoy approach, which enables the statistical determination of an overfitted result.

Email em resposta a submissão

-----Original Message----- From: eic.proteomics1@ucd.ie
Sent: Tuesday, May 31, 2011 2:57 PM
To: paulo@pcarvalho.com
Subject: PROTEOMICS - pmic.201100297
31-May-2011
Dear Dr. Carvalho,
Your manuscript entitled "Can the false-discovery rate be misleading?" has been successfully submitted online to PROTEOMICS.
Your manuscript number is pmic.201100297.
Please mention the above manuscript number in all future correspondence regarding this submission.
You can view the status of your manuscript at any time by checking your Author Center after logging into http://mc.manuscriptcentral.com/proteomics .
If you have any difficulty using the site, please contact our Support Desk at MSCentral-Support-Europe@wiley-vch.de.
Thank you for submitting your manuscript to PROTEOMICS.
Sincerely,
PROTEOMICS Editorial Office