UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

BEATRIZ JERONIMO PINTO

ProGen AP: um *pipeline* para anotação proteogenômica de *Mycobacterium tuberculosis* visando o descobrimento de genes com potencial para intervenção biotecnológica

Ribeirão Preto – SP

2012

BEATRIZ JERONIMO PINTO

ProGen AP: um *pipeline* para anotação proteogenômica de *Mycobacterium tuberculosis* visando o descobrimento de genes com potencial para intervenção biotecnológica

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para a obtenção do título de Mestre em Ciências.

Área de concentração: Genética

Orientadora: Prof^a Dr^a Silvana Giuliatti

Ribeirão Preto – SP

2012

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Pinto, Beatriz Jeronimo

ProGen AP: um *pipeline* para anotação proteogenômica de *Mycobacterium tuberculosis* visando o descobrimento de genes com potencial para intervenção biotecnológica.

Ribeirão Preto, São Paulo, 2012.

81 p. : il. ; 30 cm

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Genética.

Orientadora: Giuliatti, Silvana.

1. Proteogenômica. 2. Biomarcadores. 3. Banco de dados. 4. *Mycobacterium tuberculosis*. 5. *pipeline*.

FOLHA DE APROVAÇÃO

Beatriz Jeronimo Pinto

ProGen AP: um *pipeline* para anotação proteogenômica de *Mycobacterium tuberculosis* visando o descobrimento de genes com potencial para intervenção biotecnológica.

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para a obtenção do título de Mestre em Ciências.

Área de concentração: Genética

Orientadora: Prof^a Dr^a Silvana Giuliatti

Aprovada em:

Banca examinadora

Prof. Dr.:

Instituição:

Assinatura: _____

Prof. Dr.:

Instituição:

Assinatura: _____

Prof. Dr.:

Instituição:

Assinatura: _____

*Aos meus queridos pais, Eduardo e Maria Helena,
pois sem vocês eu jamais chegaria até aqui.*

Amo vocês.

Agradecimentos

Agradeço, especialmente, à Prof.^a Dr^a Silvana Giuliatti pela oportunidade de integração ao Grupo de Bioinformática, por todo apoio, respeito e incentivo durante a realização deste projeto e, principalmente, pelo apoio e atenção nos momentos difíceis enfrentados no decorrer do projeto.

Agradeço a disposição dos membros da banca examinadora para avaliar e discutir o presente projeto.

Agradeço ao Prof. Dr. Vitor Marcel Faça pela orientação, paciência, dedicação e atenção dispensados a minha formação e ao desenvolvimento deste projeto.

Agradeço a Dr^a Sandra Faça, coordenadora da equipe VLS, pela atenção, integração, comprometimento e pela oportunidade oferecida de desenvolver esse projeto tão interessante.

Agradeço ao Departamento de Genética e à Faculdade de Medicina de Ribeirão Preto pela estrutura oferecida para minha formação na graduação e pós-graduação.

Agradeço à FAPESP pelo apoio financeiro dado ao projeto.

Agradeço ao Departamento de Genética e ao Prof.Dr. Ademilson Espencer Egea Soares por viabilizar a participação no Curso de Proteômica oferecido no Centro de Proteômica de Seattle e que foi vital para o desenvolvimento desse projeto.

Gostaria de agradecer aos meus colegas do Grupo de Bioinformática, Iuliana, Saulo, Daniel, Nilson, Fernanda e Gabriela, pela amizade, pelo apoio e pelo convívio durante essa etapa.

Agradeço à equipe da Veritas Life Sciences (VLS), Deise, Camila, Virgínia e Sandra, pelo convívio, amizade, apoio e aprendizado durante todo o período que trabalhei com vocês.

Agradeço aos meus pais, Eduardo e Maria Helena, pela educação, pelos ensinamentos, por todo o esforço feito para que eu estivesse onde estou e por toda ajuda e carinho que sempre me deram.

Agradeço à minha irmã Marília, pela amizade, carinho, apoio e por sempre estar ao meu lado.

Enfim, agradeço a todos que de alguma forma, direta ou indiretamente, colaboraram para o meu crescimento pessoal, profissional e para a realização deste projeto. Deixo aqui o meu muito obrigado!

Apoio e Suporte Financeiro

Este projeto foi realizado com o apoio financeiro das seguintes entidades e instituições:

FAPESP

CNPQ

Faculdade de Medicina de Ribeirão Preto – FMRP/USP

Departamento de Genética

Normalização Adotada

Este documento foi elaborado de acordo com:

Referências: Associação Brasileira de Normas Técnicas (ABNT) – NBR 6023

Universidade de São Paulo. Sistema Integrado de Bibliotecas. Diretrizes para apresentações de dissertações e teses da USP: documento eletrônico e impresso. São Paulo, 2009.

RESUMO

PINTO, B.J. ProGen AP: um *pipeline* para anotação proteogenômica de *Mycobacterium tuberculosis* visando o descobrimento de genes com potencial para intervenção biotecnológica. 2012. Dissertação (Mestrado). Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo. Ribeirão Preto, 2012.

Anotação proteogenômica é uma abordagem que une a análise proteômica com a anotação genômica. O intuito de tal abordagem é prover uma anotação mais detalhada ao gene. Intuito esse, que nem sempre é possível quando se trata apenas de genes, uma vez que produtos gênicos, com funções importantes preditas, somente passam a ter papel na fisiologia do organismo quando expressos e traduzidos. Com todo o avanço atual de estudos na área proteogenômica, a geração de dados tem crescido de modo exponencial e, com esse crescimento, nota-se a necessidade cada vez maior da criação de sistemas capazes de processar, armazenar e gerenciar essas novas informações produzidas. Assim, é descrito nesse trabalho o desenvolvimento do ProGen AP, sendo constituído de uma interface *web* construída em HTML/PHP5, um banco de dados cujo SGBD é o MySQL e de módulos de processamento de dados proteômicos, neste caso o LabKey (com o core Xtandem!) e o QuickMod. Todos os módulos são *open source* e comunicam entre si através de *scripts* PERL. Nesse sistema, o pesquisador fornece dados de experimentos proteômicos e o sistema, então, os processa e retorna ao usuário informações sobre o gene expresso, a localização dos peptídeos dentro do gene aos quais pertencem e, ainda, informações quantitativas sobre o peptídeo e a proteína identificados. Além disso, o uso de um processamento esquematizado reduz a possibilidade de erro de entrada/saída de dados nos módulos intermediários do processamento. Aqui, o ProGen AP foi aplicado no estudo proteômico do *Mycobacterium tuberculosis* (MTb). Na literatura, o genoma do MTb cepa H37Rv contém apenas 4062 “open reading frames” (ORFs) preditos e o complemento funcional desse genoma, o proteoma, ainda não está totalmente elucidado. A análise do proteoma do MTb, com o uso do ProGen AP, resultou em uma lista total de 154.982 identificações de peptídeos, representando um total de 147.334 peptídeos únicos. Até o momento, foram identificadas 2.369 proteínas, cobrindo aproximadamente 58% de todo o genoma do MTB. É importante ressaltar que, dentre todas as proteínas identificadas até o momento, a maioria delas está anotada como proteínas hipotéticas em seu genoma, e, por consequência, os resultados obtidos nesse projeto confirmam e validam a existência de tais produtos gênicos. Além disso, 567 peptídeos foram identificados como N-terminal e 1229 como C-terminal, o que indica a correta predição do início e do término da tradução de tais genes. Todos esses resultados positivos confirmam que a abordagem utilizada no ProGen AP é eficiente e pode ser usada em vários outros organismos de interesse do pesquisador.

Palavras-chave: proteogenômica, biomarcadores, banco de dados, *Mycobacterium tuberculosis*, *pipeline*.

ABSTRACT

PINTO, B.J. ProGen AP: a *pipeline* for proteogenomic annotation of *Mycobacterium tuberculosis* seeking the discovery of potential genes for biotechnological intervention. 2012. Dissertation (Master). Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo. Ribeirão Preto, 2012.

Proteogenomic annotation is an approach that combines proteomic analysis and genomic annotation. The aim of this approach is to provide a more detailed annotation, which is not possible in most of the times when dealing mostly with genes, once that genomic products, with important predicted functions are only important in the organism physiology when they are expressed and translated. There have been occurring several advances in proteogenomic studies and the generation of new data sets has been growing in an exponential wave. With all this growth, the creation of systems able to storing, processing and analyzing all the new knowledge produced is eminent. This study presents the deployment of ProGen AP, a system built with a HTML/PHP5 web interface, a mySQL data management system to store the data and two processing modules (LabKey, with core X!Tandem and QuickMod). In this system, the researcher provides a data set from a proteomic experiment and then the system processes it and returns to the researcher information about the expressed gene, the peptides localization inside the gene that they belong and, also, quantitative information about the peptide and the protein that were identified. Also, the use of an automated pipeline reduces the possibility of making mistakes in input/output of the data when using the intermediate modules. Here, the ProGen AP were applied to perform a proteogenomic annotation of *Mycobacterium tuberculosis* (MTb). In literature, the MTb genome, strain H37RV, have only 4062 predicted open reading frames (ORFs) and the functional complement of this genome is not completely known. The *MTb* analysis using ProGen AP, resulted in a list of 154.982 peptides identification, representing a total of 147.334 single peptides. Until now, were identified 2.369 proteins, covering nearly of 58% of the whole MTb genome. Is very important to highlight that, among all the identified proteins until now, most of them are annotated as hypothetical proteins in the MTb genome, so can be affirmed that the results of this project can confirm and validate the existence of all these genomic products. Beside this, 567 peptides were identified as been an N-terminal peptide and 1229 were identified as been a C-terminal, this fact indicates that the prediction of the beginning and the end of translation of those genes are right. All these positive results corroborate that the approach utilized in the ProGen AP is efficient and can be used in studies of other organisms.

Keywords: proteogenomic, biomarkers, database, *Mycobacterium tuberculosis*, *pipeline*.

Lista de Figuras

Figura 1: esquema representativo dos módulos de processamento e análise dos dados.

Figura 2: esquema do fluxo de processamento das Buscas Padrão e Refinada.

Figura 3: arquivo XML contendo os parâmetros de entrada do *X!Tandem*.

Figura 4: curva de aprendizado do *PeptideProphet* (fonte: Brian Searle – *Peptide Prophet Explained*)

Figura 5: fluxograma dos módulos do TPP.

Figura 6: esquema do fluxo de processamento da Busca Genômica.

Figura 7: esquema do fluxo de processamento da Busca em *Spectral Library* (fonte: AHRNÉ, E., *et al.*, 2011).

Figura 8: Visão Geral do funcionamento do pipeline ProGen AP.

Figura 9: Relação entre as tabelas do Banco de Dados Relacional do ProGen AP.

Figura 10: Diagrama de Caso de Uso: Visão Geral dos Atores e Casos de Uso. Os atores Usuário Comum e Administrador são os usuários do sistema e interagem com os casos de uso que cada papel tem permissão para realizar. O ator Visitante não é um usuário cadastrado no sistema, podendo visualizar somente os conteúdos públicos.

Figura 11: Diagrama de Colaboração do caso de uso Pedir Cadastro. Representa todos os eventos que o ator Usuário comum realiza para solicitar que tenha login e senha cadastrados no sistema.

Figura 12: Diagrama de Colaboração do caso de uso Cadastrar Usuário. Representa todos os eventos que o ator Administrador realiza para criar um novo usuário do sistema.

Figura 13: Diagrama de Colaboração do caso de uso Busca no Banco. Representa todos os eventos que o ator Usuário realiza para fazer uma busca por peptídeo ou proteína.

Figura 14: Diagrama de Colaboração do caso de uso Inserir Dados no Banco. Representa todos os eventos que o ator Administrador realiza para poder inserir novos dados às tabelas do banco de dados do sistema.

Figura 15: Parâmetros do *X!Tandem* incluindo “*fractions*” na análise preliminar dos dados.

Figura 16: curva de sensibilidade e taxa de erro da análise preliminar dos dados.

Figura 17: resultado geral de identificação dos peptídeos.

Figura 18 Página com Introdução e Motivação do projeto.

Figura 19: Continuação da Página de Introdução, agora com resumo do projeto.

Figura 20: Página de Login/Senha.

Figura 21: Página principal de Busca.

Figura 22: Página de Visualização do Resultado Geral.

Figura 23: Página Visualização do Resultado de Busca por Peptídeo Específico.

Figura 24: Exemplo da ferramenta de visualização gráfica de alinhamentos.

Lista de Tabelas

Tabela 1: Principais modificações pós-traducionais e funções descritas (fonte MANN; JENSEN, 2003).

Tabela 2: Relação das amostras por fração. Cada xx corresponde ao número de cada amostras, podendo variar de 01 até 10.

Tabela 3: Quantidade de peptídeos identificados na análise com o parâmetro “fractions” e sem, com as amostras sendo analisadas individualmente.

Tabela 4: Relação de quantidade de identificações resultantes na Busca Padrão.

Tabela 5: Relação de quantidade de identificações resultantes na Busca Refinada.

Tabela 6: Relação de quantidade de identificações resultantes na Busca Genômica.

Lista de Abreviaturas e Siglas

2DE	Eletroforese bidimensional
2D-PAGE	two-dimensional polyacrylamide gel electrophoresis (gel bidimensional de acrilamida para eletroforese)
CPAS	Computational Proteomics Analysis System (Sistema Computacional de Análise Proteômica)
DNA	Deoxyribonucleic acid (ácido desoxiribonucléico)
EPT	Expressed Peptide Tag (Etiqueta de Peptídeo Expresso)
EST	Expressed Sequence Tag (Etiqueta de Sequência Expressa)
GPL	General Public Licence (Licença de Uso Público Geral)
GPS	General Proteomics Standard (Padronização Geral Proteômica)
HTML	Hypertext Markup Language (Linguagem de Marcação de Hipertexto)
HuPO	Human Proteome Organization (Organização do Proteoma Humano)
ISB	Institute for Systems Biology. (Instituto para Biologia de Sistemas)
LC-MS/MS	Liquid chromatography-mass spectrometry (espectrometria de massas por cromatografia líquida)
LIMS	Laboratory Information Management System
MALDI-TOF	Matrix assisted laser desorption ionization – time off light (Ionização/Dessorção de Matriz Assistida por Laser)
MIAME	Minimum Information About a Microarray Experiment (Informações Mínimas sobre um experimento de Microarranjos)
MIAPe	Minimum Information About a Proteomic Experiment (Informações Mínimas sobre um experimento Proteômico)
MS/MS	Tandem Mass Spectrometry (espectrometria de massas em série)
MTb	Mycobacterium tuberculosis
NCBI	National Center for Biotechnology Information (Centro Nacional para Informações Biotecnológicas)
mzXML	Arquivo baseado em XML específico para dados de espectrometria de massas.
NSF	National Science Foundation (Fundação Nacional de Ciências)
NHGRI	National Human Genome Research Institute (Instituto Nacional de Pesquisa do Genoma Humano)
NIGMS	National Institute of general medical Sciences (Instituto Nacional de Ciências Médicas Gerais)

NIST	National Institute of Standards and Technology (Instituto Nacional de Padronização e Tecnologia)
OMG	Object Management Group
ORF	open reading frame (quadro de leitura aberto)
PEDRo	Proteome Experimental Data Repository (Repositório de Dados Proteômicos Experimentais)
Perl	Practical Extraction and Report Language
PHP	Hypertext Pre-Processor (Pre-Processador de Hipertexto)
PTM	Post translational modifications (Modificações pós-traducionais)
QTOF	Quadrupole time-of-flight mass spectrometer (Espectrômetro de Massas Quadrupolar baseado em tempo de voo).
SGBD	Sistema gerenciador de Banco de Dados
SPC	Seattle Proteome Center (Centro de proteômica de Seattle)
SQL	Structured Query Language (Linguagem de Consulta Estruturada)
TB	Tuberculose
TOF-TOF	Tandem time-of-flight mass spectrometer (Espectrômetro de Massas em Série baseado em tempo de voo)
TPP	TransProteomicPipeline
UML	Unified Modeling Language (Linguagem de Modelagem Unificada)
VLS	Veritas Life Sciences
XML	Extensible Markup language (Linguagem Extensível de Marcação)

SUMÁRIO

1. INTRODUÇÃO	19
1.1. Genômica e Proteômica	20
1.2. Proteogenômica	21
1.3. <i>Mycobacterium tuberculosis</i> (MTb)	22
1.4. A Bioinformática e a criação de Bancos de Dados aplicados	23
2. OBJETIVOS	25
3. MATERIAIS E MÉTODOS	27
3.1. Aquisição dos Dados	28
3.2. Desenvolvimento do Pipeline	28
3.2.1. Busca Padrão e Busca Refinada	30
3.2.1.1. CPAS/LabKey	31
3.2.1.2. X!Tandem	32
3.2.1.3. PeptideProphet	33
3.2.1.4. TransProteomic Pipeline	34
3.2.2. Busca Genômica	35
3.2.3. Criação e Busca em Bibliotecas de Espectros	36
3.2.3.1. A importância das Modificações Pós-Traducionais	37
3.2.3.2. Quickmod	40
3.3. Construção do Sistema	40
3.3.1. Ferramentas Utilizadas	42
3.3.2. Banco de Dados	43
3.4. Modelagem do Sistema	46
3.4.1. Escopo do Sistema	46
3.4.2. Diagramas UML	47
4. RESULTADOS E DISCUSSÃO	53
4.1. Análise dos Resultados	54
4.2. Identificação de peptídeos	57
4.3. O sistema	60
4.3.1. Página Inicial	60
4.3.2. Área de Acesso	63
4.3.3. Pesquisa no Banco de Dados	64
4.3.4. Resultado Geral	65

4.3.5. Resultado Específico	66
5. CONCLUSÕES	69
6. REFERÊNCIAS BIBLIOGRÁFICAS	72
7. ANEXOS	77

1. INTRODUÇÃO

1.1. Genômica e Proteômica

O estudo do produto funcional da informação genética, o proteoma, é fundamental, uma vez que a estrutura, função, abundância e até o número de proteínas em um organismo não podem ser preditos apenas pela sequência do DNA. Como exemplo, podem-se citar as modificações pós-traducionais tipo fosforilação e glicosilação, que contribuem para a diversidade protéica e regulam a função de muitas proteínas (PANDEY; MANN, 2000).

A proteômica é relativamente recente como ciência e a maioria dos estudos realizados até poucos anos atrás foi desenvolvida utilizando-se a eletroforese bidimensional (2DE) e espectrometria de massas tipo “*Matrix assisted laser desorption ionization- Time of flight*” (MALDI-ToF) para separar e identificar as proteínas, respectivamente. Recentemente, a espectrometria de massas evoluiu de simples ferramenta para identificar e caracterizar proteínas isoladas ou para mapear picos de massas em amostras clínicas, para se estabelecer como uma plataforma voltada para a análise de proteomas complexos através do “matching” do espectro de massas com bancos de dados de sequências, e assim, fornecer a identificação das proteínas (COX ; MANN, 2007). Essa evolução já está tendo impacto em várias áreas da biologia, especialmente na descoberta de biomarcadores para diversas doenças, que representa uma das grandes promessas da proteômica (HANASH; PITTERI; FACA, 2008).

Entretanto, mesmo com os avanços substanciais dos espectrômetros de massas em termos de sensibilidade e exatidão na medida de massas, a complexidade do proteoma dos tecidos e fluídos biológicos excede de longe, as atuais capacidades desta técnica analítica no sentido de permitir a separação completa das proteínas individuais e peptídeos em uma única e simples análise. Portanto, estratégias atuais para obter uma cobertura extensiva do proteoma requer fracionamento da amostra seguida de análise individual das frações ou captura de subgrupos de proteínas ou peptídeos tais como proteínas glicosiladas ou fosforiladas. Essa nova geração de estudos tem resultado em um aumento substancial do número de proteínas identificadas por experimento (CRAVATT; SIMON; YATES, 2007; FACA et. al., 2007). Ainda, em combinação com marcação isotópica das amostras, milhares de proteínas podem ser rotineiramente identificadas e quantificadas simultaneamente (FACA et al, 2006).

1.2. Proteogenômica

Embora a descoberta de genes em procariotos tenha avançado significativamente, a predição de genes curtos, anotação de genes com códon não usual, bem como a predição exata de códons de inicialização, ainda representam um grande desafio. Ao mesmo tempo, o número de genomas bacteriais sequenciados aumentou significativamente, mas não a habilidade de validar os genes preditos, deixando assim uma lacuna em termos de anotação do proteoma bacterial. Jaffe, Berg e Church (2004), utilizaram dados proteômicos para a análise do genoma de bactérias relativamente pequenas, como o *Mycoplasma mobile* e teve 26 genes preditos exclusivamente por dados proteômicos (JAFJE; BERG; CHURCH, 2004). Similarmente, Gupta et al (2007), utilizaram a bactéria gram-negativa *Shewanella oneidensis* para uma anotação proteômica baseada em “*liquid chromatography-mass spectrometry*” (LC-MS/MS) (GUPTA et al., 2007). Em ambos os casos, os peptídeos identificados por espectrometria de massas (MS/MS) forneceram importantes informações a respeito de processamento pós-traducional incluindo a identificação de eventos proteolíticos. É especialmente importante enfatizar essas descobertas porque o atual conhecimento das modificações pós-traducionais em bactérias é muito limitado, mesmo para organismos bem estudados como *E. coli*.

Todas essas aplicações potenciais da proteômica, no sentido de estender a informação genômica, estão levando a um intensivo desenvolvimento de novas ferramentas para explorar mais eficientemente os dados de espectrometria de massas produzidos em larga escala (“*hight-throughput*”). Essas ferramentas permitem a utilização dos mesmos espectros adquiridos para a identificação das proteínas (através de buscas em bancos de dados convencionais) para uma adicional exploração dos dados como, por exemplo, fazendo a busca dentro dos 6 quadros de leitura (“*reading frames*”) da sequência genômica e identificando padrões de modificações de massas nos peptídeos que correspondam a modificações pós-traducionais; realizando sequenciamento *de novo* nos espectros de MS/MS, dentre outras aplicações.

Em resumo, a análise proteômica vem contribuir significativamente para a anotação mais detalhada de genomas. Produtos gênicos com funções importantes preditas somente passam a ter papel na fisiologia do organismo quando expressos. Assim a observação de proteínas, a correção de sequência de produtos gênicos ou mesmo a indicação de ocorrência de novos genes não preditos têm impacto significativo no desenvolvimento de novos alvos para diagnóstico e terapia para agentes patológicos como o *MTb*.

1.3. *Mycobacterium tuberculosis*

O bacilo da tuberculose, doença também conhecida mundialmente pela sigla TB, tem algumas características únicas que incluem um crescimento lento, dormência, envelope celular complexo, patogenia intracelular e homogeneidade genética. *MTb* é classificado como uma bactéria aeróbica e Gram-positiva de crescimento rápido em meio ácido devido a ausência de uma membrana celular externa. Sua parede celular, altamente hidrofóbica, é única entre as bactérias e é composta por longas cadeias de ácidos graxos, glicolipídeos e outros componentes. Esse robusto envelope contribui para a capacidade de *MTb* em sobreviver nos fagócitos de grande capacidade microbicida. O bacilo normalmente se divide a cada 15-20 horas e esse crescimento lento forma a base da natureza crônica da infecção e doença, complica o diagnóstico microbiológico e exige tratamento com drogas por períodos longos (KAUFMANN, 2001,2005).

O genoma completo da linhagem H37Rv de *Mycobacterium tuberculosis* contém 3.988 “open reading frames” (ORFs) preditos pela ferramenta *ORF Finder* (ROMBEL et al, 2002). Entretanto, o complemento funcional dessa informação genética, que é o proteoma, está longe de ser inteiramente elucidado. A relação estrutura-função, abundância e modificações pós-traducionais das proteínas, que não podem ser preditas somente pela sequência de DNA, são de importância fundamental para um melhor entendimento dos mecanismos patogênicos desse organismo e para o desenvolvimento de novas terapias e métodos de diagnóstico para a tuberculose (TB).

A literatura disponibiliza alguns resultados de domínio público que serviram de base para a construção de um banco de dados proteômico preliminar de *Mtb*. Alguns desses estudos estão representados no banco de dados de proteínas identificadas por eletroforese bidimensional (2DE) <http://www.mpiib-berlin.mpg.de/2D-PAGE> (MOLLENKOPF et al., 1999; MAWUENYEGA et al., 2005). Este conjunto de dados específico contém apenas 231 proteínas celulares e 138 proteínas constituintes do sobrenadante da cultura de linhagem H37Rv. Em contrapartida, dados gerados por cromatografia líquida acoplada a espectrometria de massas de alta eficiência (LC-MS/MS) permitiram uma cobertura do proteoma de maneira significativamente melhor. Um único trabalho contribuiu com a maioria (1.044) de todas as proteínas já observadas em *MTb* (MAWUENYEGA et al., 2005).

Todos os dados proteômicos publicados até o início do presente projeto suportavam a expressão de 1.586 proteínas, ou seja 40% de cobertura dos 3988 genes codificadores de proteínas. A identificação de menos da metade dos genes preditos em *MTb* em mais de 10

trabalhos independentes indicam que as técnicas analíticas utilizadas, na sua maioria, não foram suficientes para resolver a desafiadora complexidade do proteoma deste bacilo.

1.4. A Bioinformática e a Criação de Bancos de Dados Aplicados

Com todo esse avanço dos estudos na área proteogenômica, a geração de dados vem crescendo de modo exponencial e, com esse crescimento, nota-se a necessidade cada vez maior da criação de repositórios para armazenar e gerenciar essas novas informações produzidas. Muitos repositórios de dados foram construídos para armazenar esse tipo de informação, como, por exemplo, SWISS-2DPAGE, o *PeptideAtlas* e o NIST *Libraries of Peptide Tandem Mass Spectra*.

O SWISS-2DPAGE (HOOGLAND et al., 2004) é um repositório de experimentos de eletroforese em géis bidimensionais (2-D PAGE), foi criado em 1993 e contém dados de identificação de proteínas em vários organismos. Atualmente, o SWISS-2DPAGE está sob os cuidados do grupo Swiss-Prot e da estrutura do Vital-IT do Instituto Suíço de Bioinformática. O SWISS-2DPAGE pode ser acessado pelo link <http://world-2dpage.expasy.org/swiss-2dpage/>.

O repositório *PeptideAtlas* (DESIERE., et al, 2004) é um compêndio de peptídeos provenientes de vários organismos e que foram identificados em um grande grupo de experimentos de espectrometria de massas em série (*tandem mass spectrometry* – MS/MS). Todos os dados obtidos pelos experimentos de espectrometria desse repositório foram processados pelo TransProteomic Pipeline (DEUTSCH et al, 2010) para garantir uma uniformidade nas identificações, reprodutibilidade dos experimentos e acurácia tanto na descoberta de verdadeiras identificações de peptídeos quanto nas taxas de falsos positivos. Esse repositório disponibiliza ainda os dados brutos, sem nenhuma análise, o que possibilita outros estudos por diferentes pesquisadores. Tanto o repositório quanto o sistema de processamento são mantidos pelo SPC - *Seattle Proteome Center*, um grupo-multidisciplinar, mantido por várias instituições de pesquisa (*National Institute of General Medical Sciences* – NIGMS, *National Science Foundation* – NSF, *National Human Genome Research Institute* – NHGRI, além do Ducado de Luxemburgo) e com a sede no *Institute for Systems Biology* (ISB), em Seattle-WA.

Diferentemente dos anteriores, o NIST *Libraries of Peptide Tandem Mass Spectra* (STEIN; RUDNICK, 2004) é um repositório onde o principal dado armazenado é o espectro obtido no experimento, em vez da identificação correlata do peptídeo. Esse repositório foi

desenvolvido pelo *Chemical Reference Data Group*, um grupo dentro do *National Institute of Standards and Technology*. Esse repositório possui dados de experimentos de espectrometria de massas (MS/MS) provenientes de diversos tipos de espectrômetros e também de diversos organismos, tendo um número significativamente maior de espectros de experimentos de *H.sapiens*.

Todos os repositórios acima citados são sistemas bastante reconhecidos, com dados muito confiáveis e utilizados em muitos outros estudos e pesquisas proteômicas. É importante ressaltar que todos esses repositórios citados, possuem uma diversidade de plataformas de coletas e análises de dados, cada uma com seu diferencial, podendo cada uma delas gerar identificações um pouco diferentes a partir de uma mesma amostra. Isso é devido, ainda, a uma multiplicidade de métodos de quantificação dos peptídeos obtidos.

Uma deficiência desses repositórios, porém, é a pouca disponibilização pública de ferramentas que integrem o armazenamento, processamento e distribuição de dados proteômicos em larga escala. A integração entre sistemas, juntamente com a re-análise de vários conjuntos de dados provenientes de diferentes experimentos proteômicos, são necessários durante o estudo exploratório de funções de proteínas em alguns processos biológicos, como no projeto aqui descrito.

Para alcançar tais propósitos, os sistemas de repositório de dados precisam cumprir alguns critérios específicos durante a troca e o armazenamento dos dados. Geralmente, utiliza-se a linguagem *Extensible Markup Language* (XML) para a especificação de critérios a respeito de dados biológicos (TAYLOR et al., 2007; QUACKENBUSH, 2006). Um exemplo amplamente difundido é o documento constando as informações mínimas sobre um experimento de *microarrays*, *Minimum Information About a Microarray Experimente* (MIAME). O MIAME é um documento que descreve quais são as informações básicas necessárias para se descrever um experimento de expressão gênica (BRAZMA et al., 2001) e se tornou mundialmente aceito como um protocolo padrão para a realização de estudos com dados de *microarrays* (PARKINSON et al., 2007). Com esse mesmo intuito, nasce a normativa MIAPE (TAYLOR et al., 2007). O objetivo dessa normativa é incentivar a criação, integração, armazenamento e disseminação de dados de experimentos proteômicos e, para tanto, foram desenvolvidos módulos para a descrição do uso das mais diversas técnicas dessa área como, por exemplo, a utilização de géis de eletroforese, a espectrometria de massas e outras. Assim, baseando-se na normativa MIAME, foi construído, no presente projeto, um sistema de banco de dados que permite a captura, anotação e disseminação de dados e informações geradas nesse estudo do organismo *M. tuberculosis*.

2. OBJETIVOS

O objetivo geral deste projeto é a criação de uma infraestrutura de Tecnologia da Informação, incluindo o desenvolvimento de novos algoritmos e/ou a re-utilização e adaptação de componentes de código “open source” já existentes, para a montagem de um sistema *in silico* (ProGen AP) capaz de integrar dados proteômicos de alta confiabilidade (*Expressed Protein Tags*, EPTs) (GUPTA et al., 2008) com a informação genômica disponível (e.g., genomas completos, mapas genômicos, bases de dados de ESTs) para organismos importantes.

Metas Específicas

Para atingir o objetivo acima proposto, as metas específicas a serem executadas serão (em ordem cronológica):

- a) Obtenção de “datasets” modelos de EPTs de *Mtb* de alta qualidade a partir da análise proteômica de ampla cobertura de cepas derivadas da *MTb* H37Rv cultivadas em condições aeróbicas e anaeróbicas.
- b) Criação de um banco de dados relacional para a captura, armazenamento, “mining”, distribuição e anotação dos dados proteômicos (i.e., “raw data”, metadata e EPTs).
- c) Desenvolvimento e/ou integração com o sistema de banco de dados de módulos de mapeamento e localização dos EPTs dentro de (os) genoma(s) de referência (e.g., *Mtb* H37Rv).
- d) Desenvolvimento e/ou integração com o banco de dados de módulos de aglomeração espectral (“*spectral clustering*”) (FALKNER et al., 2008) que maximizem a cobertura proteômica (porcentagem de EPTs identificados e mapeados, incluindo suas modificações pós-traducionais.)
- e) Criação de um portal virtual de acesso controlado ao genoma anotado do *MTb* para a comunidade científica.

**MATERIAIS E
MÉTODOS**

3. MATERIAIS E MÉTODOS

3.1. Aquisição dos Dados

Os experimentos realizados para a obtenção dos espectros, que são o alvo de estudo desse trabalho, foram feitos pela equipe da empresa Veritas Life Sciences (VLS), sob coordenação da pesquisadora Sandra Faça. A abordagem de fracionamento de proteínas utilizada para o sequenciamento proteômico deste projeto gerou um total de 180 frações do proteoma do *MTb*. Cada uma destas frações foi analisada individualmente por espectrometria de massas com cromatografia líquida (LC-MS/MS), o que representou um total de, aproximadamente, 520 horas de aquisição de dados e resultou em 1.800.000 espectros. O volume de dados gerado neste projeto é comparável ao obtido no sequenciamento de genomas, resultando em longos tempos de processamento e análise dos resultados. Neste projeto, estes dados foram processados várias vezes por diferentes abordagens, destacando a infra-estrutura computacional como o maior gargalo do processo, necessitando aproximadamente 20 dias de processamento ininterrupto e dedicado de um computador Dell Vostro 220s, com processador core 2duo e memória RAM de 4Gb, 800Mhz para a execução do *pipeline* inteiro, depois de estabelecido e testado cada módulo do sistema.

3.2. Desenvolvimento do *pipeline*

A construção de *pipelines* para a anotação proteogenômica “*high-throughput*” e automática representa uma área de pesquisa relativamente nova e de muita atividade dentro do campo da bioinformática atual. O *pipeline* desenvolvido nesse projeto buscou estabelecer uma sequência de procedimentos para obtenção de informações sobre anotação proteogenômica de acesso controlado, bem como sua aplicação à anotação funcional comparativa do genoma de *Mycobacterium tuberculosis*.

Durante a fase de execução do projeto e montagem do *pipeline*, utilizou-se ferramentas as quais pudessem ser aplicadas posteriormente em outros estudos, com alvos biológicos diferentes, mas que tivessem abordagem experimental semelhante ao projeto atual. Dessa maneira, garante-se a criação de um sistema/*pipeline* que fosse geral para estudos de análises proteogenômicas, independentes dos alvos terapêuticos, poupando tempo na fase de desenvolvimento, e garantindo um tempo maior para a análise dos resultados.

Devido ao grande volume de dados experimentais adquirido e ao grande tempo computacional que demandaria processar todos os dados, decidiu-se, inicialmente, percorrer o

pipeline até o final com apenas uma fração dos dados, analisar os resultados obtidos e, somente depois de todos os passos definidos, submeter todas as frações para análise. Essa estratégia garantiu uma maior rapidez na definição das ferramentas que iriam constituir cada módulo do *pipeline*.

Os dados gerados para as diferentes frações foram combinados e processados de 4 maneiras diferentes: i) busca em bancos de dados de proteínas preditas por peptídeos trípticos (denominada aqui Busca Padrão); ii) busca em bancos de dados de proteínas preditas por peptídeos semi-trípticos (Busca Refinada); iii) busca em banco de dados genômico e iv) busca por peptídeos em biblioteca de espectros (*spectral library search*).

É via de regra, quando em experimentos de espectrometria de massas, o uso de algoritmo de busca e identificação de peptídeos. Geralmente, esses algoritmos são pagos e distribuídos com o equipamento, permitindo ao pesquisador a instalação em algumas máquinas, dependendo da quantidade de licenças distribuídas com o mesmo. Dentre esses *softwares* (Anexo A) existem o Sequest (ENG; McCORMACK; YATES, 1994) distribuído pela ThermoScientific, o Mascot (Matrix Science) (PERKINS et al., 1999), Xcalibur (Thermo Fisher Scientific) e o PEAKS DB (Bioinformatics Solution Inc.) (Zhang et al., 2012), além do OMSSA (GEER et al, 2004), desenvolvido pelo NCBI. O procedimento comum em experimentos proteômicos é que, após a amostra ser processada pelo espectrômetro de massas, os dados brutos sejam analisados nesses *softwares* e os resultados obtidos, então sejam utilizados como resultados finais. Porém, quando se utiliza apenas esse procedimento comum, o volume de dados obtidos é bem menor do que se fosse feita uma análise paralela com diferentes algoritmos, nos quais, cada um deles tivesse, como ponto forte, uma característica importante no estudo de proteínas: identificação de peptídeos comuns, identificação de peptídeos que apresentem alguma modificação prevista, identificação de peptídeos considerados “não ideais” e identificação de peptídeos por metodologia *denovo*.

Conforme destacado anteriormente, esta abordagem paralela foi adotada no presente projeto, visa a identificação do maior número possível de peptídeos do *MTb* e está representada graficamente na Figura 1. Destacam-se ainda que todos os resultados de identificação apresentados aqui correspondem a uma primeira rodada de análises, onde altos critérios de qualidade de identificação foram utilizados a fim de minimizar falso-positivos. Todos os dados serão continuamente refinados a medida que novas ferramentas sejam implementadas, ou mesmo que dados adicionais sejam gerados ou disponibilizados por outros grupos.

Figura 1: Esquema representativo dos módulos de processamento e análise dos dados.

3.2.1. Busca Padrão e Busca Refinada

O processo de identificação de peptídeos nos dados de espectrometria de massa iniciou-se pelo método padrão de identificação, onde os espectros são correlacionados com espectros simulados para peptídeos trípticos presentes em um banco de dados de proteína. Esta análise foi realizada utilizando o algoritmo X!Tandem (CRAIG; BEAVIS, 2004), (algoritmo “open source” incorporado no sistema *Computational Proteomics Analysis System* - CPAS (RAUCH A, et al., 2006) e parte integrante do software LabKey (NELSON, et al., 2011) e o banco de dados de proteínas preditas no genoma da cepa H37Rv do *MTb*, disponível no GeneBank (VERSION NC_000962.2 de 09/Nov/2009). Apesar de poderem gerar mais identificações, não foram considerados nesta primeira rodada de análises, que denominamos de Busca Padrão, os peptídeos semi-trípticos, que apresentam sítio para a

tripsina em apenas uma das extremidades, uma vez que o tempo de processamento destes resultados é significativamente maior.

Os peptídeos semi-trípticos foram considerados em um segundo momento, o qual foi denominado Busca Refinada. Nessa busca, o algoritmo faz uma digestão *in silico* semi-tríptica, na qual o algoritmo assume que um dos sítios de clivagem é triptico ([Kr]{p}), mas o outro sítio pode ser qualquer um dos resíduos. Essa abordagem gera um número muito maior de identificação de peptídeos, principalmente peptídeos mais curtos, o que aumenta proporcionalmente o tempo de processamento dos dados, bem como o volume de dados resultante, uma vez que o campo de busca de espectros aumentou, pelo menos, três vezes, devido às várias possibilidades de combinação de identificação de um mesmo peptídeo. O funcionamento da Busca Padrão e da Busca Refinada são ilustrados na Figura 2.

Figura 2: Esquema do fluxo de processamento das Buscas Padrão e Refinada.

3.2.1.1. CPAS/ LabKey

O CPAS, sigla para Sistema Computacional de Análise Proteômica (*Computational Proteomics Analysis System*), é a ferramenta de análise proteômica que integra a plataforma **LabKey** (NELSON et al, 2011). O CPAS é uma ferramenta integrante do software Labkey (NELSON et al., 2011), que é um *laboratory information management system* (LIMS) e funciona especificamente para laboratórios de pesquisa focada em processos e experimentos proteômicos. Essa ferramenta utiliza módulos *open-source* criados e disponibilizados pelo TransProteomicPipeline (DEUTSCH et al., 2010), desenvolvidos no *Institute for Systems Biology* (ISB). O CPAS também é, em si, uma ferramenta *open-source* e contém um *pipeline* completo para análise e administração de dados provenientes de experimentos proteômicos de Espectrometria de Massas por Cromatografia Líquida (LC-MS/MS), incluindo anotação proteômica, busca em bancos de dados protéicos e busca por peptídeos em espectros e identificação de proteínas (RAUCH et al., 2006).

O sistema importa e processa dados MS/MS provenientes de arquivos brutos e arquivos no formato mzXML (padrão). O passo seguinte é a busca nos arquivos de dados por peptídeos com *score* significativo, utilizando o mecanismo de busca do algoritmo X!Tandem os arquivos mzXML e o banco de dados fornecido no formato FASTA. Uma vez que os dados foram processados, buscados e pontuados por *score* (todas essas etapas com o uso do X!Tandem), o sistema, em seguida, executa o PeptideProphet (KELLER et al. 2002), que faz a análise quantitativa e estatística em cima dos resultados obtidos.

3.2.1.2. X!Tandem

O X!Tandem é um algoritmo de busca de proteínas, é *open source* e tem a habilidade de busca automática por sítios de clivagem, peptídeos semi-trípticos, algumas modificações pos-traducionais previamente informadas e algumas mutações pontuais (CRAIG; BEAVIS, 2003). O algoritmo X!Tandem é uma boa alternativa em relação aos mecanismos de busca mais tradicionais como SEQUEST e Mascot, pois possui uma interface de programação (API) muito simples: precisa apenas de um arquivo de entrada de comandos em XML, mostrado na Figura 3, e tem como formato de saída também um arquivo XML.

O algoritmo de busca funciona comparando cada um dos espectros em série, obtidos dos experimentos, com um espectro-modelo criado para cada entrada do banco de dados de proteína. Diferentemente dos mecanismos de busca anteriores ao X!Tandem, esse algoritmo calcula a confiança (valor esperado) para todas as atribuições de sequência ao espectro. Além

disso, remonta todas as associações dos peptídeos identificados em um conjunto de dados referente às sequências de proteínas e, a partir dessa remontagem, atribui uma confiança estatística de que a identificação e o alinhamento não são ao acaso.

Além de ser o algoritmo principal de processamento e identificação de peptídeos constituinte do LabKey (NELSON et al., 2011), o X!Tandem possui uma interface de programação muito fácil de usar e seus parâmetros são facilmente modificáveis, sendo editados em um arquivo XML, como ilustrados na Figura 3. Outro fator importante é a possibilidade de se mudar a ferramenta de busca (CPAS/LabKey), sem mudar o algoritmo de busca, uma vez que o X!Tandem é parte integrante de grande parte dos programas de identificação de proteínas, como o TPP (TransProteomePipeline) que também foi testado nesse projeto.

```

1 <bioml>
2   <note label="residue, potential modification mass" type="input">242.101505@K,3.0188302@C,15.994915@M</note>
3   <note label="residue, modification mass" type="input">71.037114@C</note>
4   <note label="protein, cleavage site" type="input">[KR]||[P]</note>
5   <note label="scoring, maximum missed cleavage sites" type="input">2</note>
6   <note label="spectrum, use conditioning" type="input">no</note>
7   <note label="scoring, minimum ion count" type="input">1</note>
8   <note label="pipeline, database" type="input">Mycobacterium_tuberculosis_H37Rv.fasta</note>
9 </bioml>

```

Figura 3: Arquivo XML contendo os parâmetros de entrada do X!Tandem

3.2.1.3. PeptideProphet

O PeptideProphet é o algoritmo utilizado para validar as atribuições de peptídeos para cada espectro de MS/MS identificado por uma das ferramentas de busca em bancos de dados (X!Tandem, SEQUEST, Mascot e outros). É uma parte imprescindível do *pipeline*, pois é ele quem determina quais identificações são válidas ou não (questão chave quando trata-se de diminuir a quantidade de falsas identificações positivas). O algoritmo utiliza estatística bayesiana para converter os *scores* em probabilidades. O método aprende a distinguir atribuições de peptídeos corretas de incorretas dentro de cada um dos conjuntos de dados e processa, para cada determinação de peptídeo, qual a probabilidade dessa atribuição estar correta (Figura 4). Foi escolhido para ser utilizado nesse projeto, pois possui a capacidade de normalizar e comparar/agregar dados provenientes de qualquer equipamento (MALDI-TOF,

TOF-TOF, QTOF, Orbitrap, etc) e de qualquer módulo de busca de peptídeos, que possuem métodos de *scoring* diferentes.

Figura 4: Curva de aprendizado do PeptideProphet (fonte: Brian Searle – Peptide Prophet Explained)

Popularmente conhecido como TPP, o TransProteomic Pipeline (DEUTSCH et al., 2010) é uma ferramenta gratuita para análises proteômicas de experimentos de MS/MS desenvolvida no *Institute for Systems Biology*. O TPP é um sistema que integra várias ferramentas distintas (algumas delas utilizadas também neste projeto) sob uma interface gráfica que pode ser utilizada pelo usuário final através de um navegador. Dentre as ferramentas integradas por esse sistema podem ser citadas: MASCOT, *X!Tandem*, *PeptideProphet*, *ProteinProphet* (NESVIZHISKII et al., 2003), XPRESS (HAN et al., 2001) e outras, como pode ser visto na Figura 5.

Figura 5: Fluxograma dos módulos do TPP

Pode-se afirmar que o TPP é o modelo de *pipeline* utilizado para a construção do *pipeline* apresentada nesse trabalho. A intenção é utilizar o fluxo do *pipeline* demonstrado no TPP, porém com outras ferramentas e utilizando alguns dos módulos, visando obter a maior quantidade de espectros identificados com máxima acurácia possível.

3.2.2. Busca Genômica

A Busca Genômica é o modulo seguinte à busca em bancos de dados de informações proteômicas. Nessa busca, o diferencial é a utilização de informações do genoma do *MTb* em vez do uso direto de bancos de dados de informações proteômicas. A razão pela qual utilizou-se essa estratégia é a possibilidade de capturar informações adicionais que ainda não foram depositadas em bancos de dados de proteínas. Uma vez que o genoma está totalmente disponível, pode-se utilizar de tal informação para corroborar e enriquecer o campo de busca.

Para construir o banco de dados genômico, obteve-se o genoma completo do *MTb* cepa H37Rv, disponível no repositório do NCBI. O genoma completo tem tamanho de 4,41Mb e é obtido em formato FASTA contendo todo material genético em bases nitrogenadas (A,T,C e G). Depois de obtido o genoma, o passo seguinte foi traduzi-lo em todos os 6 *reading-frames*. Ou seja, para cobrir todas as possibilidades do processo biológico de tradução, foi feita a tradução de três em três bases para o respectivo aminoácido ate o final, seis vezes seguidas, diferenciando cada tradução na base inicial da tradução. Na primeira tradução utilizou-se todas as n bases do genoma, combinadas três a três. Na segunda tradução, utilizou-se n-1 bases, pulando a primeira base do genoma, e traduzindo as demais, três a três e assim por diante ate iniciar na base (n-5).

Em seguida, o genoma traduzido foi separado em fragmentos contendo 200 aminoácidos, mas que fossem sobrepostos uns aos outros em 50 aminoácidos de cada lado (inicio e fim), tendo assim 50 bases intercaladas entre o fragmento anterior, o fragmento atual e o próximo fragmento. Os fragmentos foram, então, colocados em um arquivo FASTA devidamente identificados pelo seu número sequencial e o *n-frame* a que pertence (exemplificado no ANEXO B).

Após a busca, os peptídeos inéditos identificados pelo módulo, que também foram processado pelo CPAS/LabKey com core X!TANDEM, foram relacionados com os genes através da posição de cada fragmento no genoma, pois cada gene também possui um identificador com sua posição de inicio e fim no genoma.

Além de ser útil para remontar as informações obtidas nessa busca, as informações de posição do gene no genoma também foi útil para a montagem da ferramenta de alinhamento, na interface de visualização dos resultados. Todo o fluxo de informação dessa etapa está representado na Figura 6.

Figura 6: Esquema do fluxo de processamento da Busca Genômica

3.2.3. Criação e Busca em Bibliotecas de Espectros

Experimentos de espectrometria de massa em série (MS/MS) geralmente produzem conjuntos de dados redundantes, contendo vários espectros provenientes do mesmo peptídeo (FRANK et al., 2007). Uma forma de diminuir o espaço de busca a ser percorrido pelo algoritmo é diminuindo o número de espectros a serem identificados, uma vez que cada peptídeo é frequentemente identificado em vários espectros diferentes.

Na tentativa de se melhorar a porcentagem de espectros identificados, bem como a qualidade dos *scores* de tais espectros, utilizou-se a abordagem de criar uma Biblioteca de Espectros (*Spectral Library*). A busca em bibliotecas de espectros é um método que compara espectros de massa que ainda não foram identificados com uma coleção de espectros de referência, selecionados previamente (FREWEN; MACCOSS, 2007). Nesse método, uma coleção de espectros que já foram identificados em etapas anteriores do *pipeline* (Busca Padrão e Busca Refinada) e que possuem um bom *score* de identificação (PepProhet maior ou igual a 0.7) é utilizada como a biblioteca de referência. Um novo espectro deve ser identificado se for suficientemente similar a algum, ou alguns, espectro(s) presente(s) na biblioteca.

É uma etapa importante no processamento, pois nessa etapa é possível identificar alguns peptídeos com *scores* altos, que não foram identificados pelos módulos anteriores pois possuem, por exemplo, alguma modificação pós-traducional em sua constituição.

A estratégia utilizada para a busca está representada na Figura 7.

Figura 7: esquema do fluxo de processamento da Busca em *Spectral Library*
(fonte: AHRNÉ E, et al,2011)

1º passo) Os dados do experimento são analisados com o *X!Tandem* (podendo haver mais de um algoritmo de busca padrão)

2º passo) Identificações válidas de peptídeos são utilizadas para compilar uma biblioteca de espectros. O *software DeLiberator* é utilizado para criar uma biblioteca complementar, contendo espectros aleatórios (*decoy*).

3º passo) Os espectros de MS/MS que não foram identificados no primeiro passo são submetidos ao QuickMod para identificação de peptídeos modificados.

3.2.3.1. A Importância das Modificações Pós-Traducionais

A identificação de modificações proteicas é algo de vital importância para entender a regulação e a dinâmica de um proteoma. Existe um grande interesse em um complexo conjunto de modificações pós-traducionais (*post translational modifications* – PTM) com significados biológicos, como por exemplo as fosforilações, acetilações, glicosilações e metilações (AHRNÉ et al, 2011). Existem inúmeros estudos na literatura sobre PTM em experimentos de MS/MS com resultados bastante relevantes. Ficarro et al. (2002), Steen et

al. (2002) e Bausoleil et al. (2002), conseguiram demonstrar que proteínas fosforiladas exercem um importante papel em redes de sinalização. Dessa mesma forma, Tissot et al. (2009), conseguiram desvendar, através de experimentos de MS/MS, o papel da glicosilação como uma modulação de função em casos de produtos secretados e/ou proteínas de membrana. A Tabela 1 mostra as principais modificações pós-traducionais presentes em experimentos proteômicos.

Existem algumas limitações ao se tentar detectar modificações pós-traducionais utilizando MS/MS. Primeiramente, algumas PTMs só estão presentes em pequenas concentrações e, é possível, que o espectrômetro de massas falhe ao selecionar esses peptídeos para serem fragmentados. Além disso, um experimento de proteômica por *shotgun* produz um volume muito grande de dados para se ter cada espectro analisado e interpretado manualmente. Sem contar a possibilidade de haver modificações nos peptídeos durante o preparo da amostra, ou mesmo a possibilidade de haver mais de uma PTM em um peptídeo. Sendo assim, o estudo de modificações proteicas exige um desenho experimental bem cuidadoso, que envolve protocolos de fracionamento e enriquecimento de amostras para a detecção de proteínas pouco abundantes (AHRNÉ et al, 2011) e para evitar possíveis mudanças dos peptídeos durante o experimento. Os dados de MS precisam ser muito acurados e, principalmente, necessita-se de ferramentas capazes de analisar dados de MS/MS exaustivamente em busca das várias modificações, o que demanda um tempo e recursos computacionais razoáveis. A presença de modificações complica bastante a análise de bioinformática, isso se deve principalmente ao alto crescimento de possíveis candidatos de peptídeos (agora com cada uma das modificações) para espectro não identificado.

Em resumo, a etapa de busca em biblioteca de espectros é importante nesse estudo pois possibilita que sejam identificados espectros com ótima qualidade. Espectros tais que diferem dos demais, previamente identificados, apenas por uma diferença de massa decorrente de uma modificação ocorrida após a tradução.

Tabela 1: Principais modificações pos-traducionais e funções descritas (fonte: modificado de MANN; JENSEN, 2003)

Tipo de modificação (PTM)	Δ Mass (Da) ^A	Estabilidade ^B	Funções e observações
Fosforilação			Reversível, ativação/inativação de atividade enzimática, sinalização de vias
- pTyr	+80	+++	
- pSer, pThr	+80	+ / ++	
Acetilação	+42	+++	Estabilidade proteica, proteção de cauda N-terminal, histonas
Metilação	+14	+++	Regulação de expressão gênica
Glicosilação			Proteínas secretadas, reconhecimento célula-célula, regulação, reversível
- N-linked	>800	+ / ++	
- O-linked	203, >800	+ / ++	
Âncora GPI	>1.000	++	Âncora licosilfosfatidilinositol (GPI), ligante de membrana entre enzimas e receptores
Hidroxiprolina	+16	+++	Estabilidade proteica e interação proteína-ligante.
Sulfação (sTyr)	+80	+	Modulador de interação proteína-proteína e receptor-ligante.
Formação de ponte bissulfito	-2	++	Estabilidade proteica, ligações cruzadas intra e intermolecular.
Deamidação	+1	+++	Possível regulador de interações proteína-ligante e proteína-proteína, artefato químico comum.
Ácido piroglutâmico	-17	+++	Estabilidade proteica, bloqueio da cauda N-terminal
Ubiquitinação	>1.000	+ / ++	Sinal de destruição. Depois da digestão triptica, o sítio de ubiquitinação é modificado com o dipeptídeo Gly-Gly
Nitratação da tirosina	+45	+ / ++	Dano oxidativo que ocorre durante inflamação.

^A Uma lista maior de valores de Δ Mass de PTM pode ser vista em <http://abrf.org/index.cfm/dm.home>

^B Estabilidade: + lábil em MS/MS, ++ moderadamente estável, +++ estável

3.2.3.2. QuickMod

QuickMod é um módulo de busca em Bibliotecas de Espectros baseados em dados de MS/MS e projetado para identificar peptídeos modificados após a tradução (AHRNÉ et al, 2011). Como dito anteriormente, uma biblioteca de espectros é um conjunto de dados de experimentos de espectrometria de massa cujos espectros já foram identificados corretamente. Geralmente, um peptídeo que está presente na biblioteca é representado por um consenso de espectros, ou seja, um conjunto de espectros proveniente do mesmo experimento e que identificam o mesmo peptídeo.

O que o algoritmo do QuickMod basicamente faz é assumir que a diferença de massa que existe entre o espectro que está sendo identificado e um alvo presente na biblioteca pode ser explicada por uma modificação pos-traducional. Baseado nessa premissa, os dois espectros são, então, alinhados e, a partir desse momento, uma lista de medidas de similaridades é calculada. O esquema de avaliação e pontuação do QuickMod foi desenvolvido através de observação de várias combinações de intensidade de picos e da cobertura de todos os picos dos espectros analisados e dos modelos na biblioteca (KÄLL et al., 2007).

Uma ressalva importante para que o *software* fosse escolhido para ser parte desse módulo é que o QuickMod fornece meios eficientes de avaliar a qualidade do conjunto de dados de MS/MS prevendo modificações que foram induzidas na amostra pela manipulação e preparação dos experimentos (como, por exemplo, a oxidação). Porém, mais importante é que o algoritmo é capaz de identificar modificações pós-traducionais conhecidas, quando é fornecido a ele uma lista de modificações prévias e já detectadas.

3.3. Construção do Sistema

Dado o objetivo de se criar e organizar uma base de dados relacional que permitisse a captura, anotação e disseminação de dados e informações proteogenômicas, percebeu-se, durante o estudo do projeto, a existência de uma padronização tanto do formato no qual os dados são disponibilizados quanto na arquitetura do sistema a ser construído. Tais padronizações foram explanadas na fase de criação de projeto do sistema, garantindo, dessa maneira, que tanto a estrutura como o conteúdo desta base de dados fossem organizados desde o início de acordo com os padrões modernos estabelecidos pela comunidade científica internacional.

O sistema de banco de dados foi construído analisando alguns pontos principais, entre eles o fluxo de saída de informações da análise por espectrometria de massas, as convenções

contidas nas normativas MIAPE (“*Minimal Information About Proteomics Experiments*”) (TAYLOR et al., 2007; QUACKENBUSH, 2006), o acesso restrito e o tipo de informação que se desejava extrair através da análise dos dados (*data mining*). A arquitetura do repositório de dados foi baseada no *schema* do *Proteome Experimental Data Repository* (PEDRo) que está de acordo com os padrões do “*General Proteomics Standard*” (GPS) formalizado pelo “*Proteomics Standards Committee*” da “*Human Proteome Organization*” (HuPO).

O sistema foi, ainda, projetado a partir de modelos de sistemas já existentes, pesquisados em artigos e em publicações na área proteogenômica. Em 2003, Taylor e colaboradores fizeram uso da linguagem UML (*Unified Modeling Language*) para esboçar um Repositório de Dados de Experimentos Proteômicos (*Proteome Experimental Data Repository* – PEDRo) (TAYLOR et al., 2003). Desde então, PEDRo vem sendo o modelo de sistema mais utilizado para se traçar os requisitos de repositórios de dados proteogenômicos. PEDRo é um aplicativo em Java que gera, dinamicamente, formulários para modelos de dados que estão escritos em um formato de *schema* XML, produzindo, dessa maneira, arquivos de dados XML que podem ser utilizados com tal *schema*. O programa usa um sistema de navegação intuitivo baseado em árvores (*tree-based*), possui uma interface sofisticada para a povoação com dados nos campos dos formulários com termos provenientes de um vocabulário controlado. O programa tem, ainda, a habilidade de importar dados/experimentos a partir de arquivos tabulares (*tab delimited text files*) e implementa ainda várias rotinas de validação (GARWOOD et al., 2004).

Levando-se todos os fatores principais em consideração, o sistema foi projetado e uma visão geral do fluxo de seu funcionamento está representado na Figura 8. Os dados resultantes do experimento de espectrometria de massas em série (MS/MS) passam por uma mudança de formato de dado bruto (.raw) para o formato de arquivo padrão para experimentos de MS/MS (.mzXML). Já nesse formato, os dados passam a ser processados pelo *pipeline*, representado na figura 1. Depois de ter o máximo de peptídeos identificados, o usuário Administrador insere todas as identificações no Banco de Dados do sistema. Do outro lado da interface, o usuário comum acessa o sistema mediante uma interface *web*, efetuando uma busca por uma proteína/peptídeo de interesse (QUERY) e, após o sistema efetuar a busca no Banco de Dados, mostra as informações nominais e gráficas, com o uso da ferramenta de alinhamento e informações do Banco de Dados Genômico do sistema.

Figura 8: Visão geral do funcionamento do pipeline ProGen AP.

3.3.1. Ferramentas Utilizadas

Para a construção do repositório de dados foi desenvolvida uma base de dados relacional e uma interface que permite aos pesquisadores realizarem pesquisas via *web*. Foram utilizados para programação o HTML e as linguagens Linguagem de Consulta Estruturada (*Structured Query Language*, SQL), PHP e Perl.

O SQL foi desenvolvido para ser uma linguagem padrão para operações com Banco de Dados. A linguagem SQL, é uma linguagem de pesquisa declarativa para banco de dados relacional e é independente de *hardware* ou de *software*. Ao usar o SQL, não é necessário saber a respeito do *software* de banco de dados ou do *hardware* envolvido em uma operação (OLIVEIRA, 2002). O PHP, juntamente com o HTML (Hypertext Markup Language), são linguagens de programação e de marcação usadas para criação de páginas *web*. (CONVERSE; PARK, 2000).

Perl (*Practical Extraction and Report Language*) é uma linguagem de programação estável, multiplataforma e livre (GPL). É uma linguagem de programação interpretada e de alto nível, podendo ser utilizada de forma procedural ou orientada a objeto. Foi desenvolvida para ser prática e eficiente na manipulação de texto, processamento de cadeias (*strings*) e

expressões regulares (*pattern matching*), o que permite tempo de desenvolvimento curto. BioPerl é uma coleção de módulos escritos em Perl que facilita o desenvolvimento de *scripts* relacionados a aplicativos de Bioinformática. É um projeto *open source* mantida pela “*Open Bioinformatics Foundation*” e desempenhou papel fundamental no Projeto Genoma Humano (STEIN, 1996). O uso de BioPerl permitiu o desenvolvimento de módulos de programas para muitas das etapas básicas de programação na Bioinformática, entre elas:

- Acesso a sequências de nucleotídeos e peptídeos a partir de bancos de dados locais e remotos;
- Transformação dos formatos de bancos/arquivos de dados;
- Manipulação de sequências individuais;
- Criação e manipulação de alinhamentos de sequências;
- Busca por genes e outras estruturas no DNA;
- Desenvolvimento de anotações lidas automaticamente por algoritmos de busca.

Na construção da base de dados foi utilizado o Banco de Dados MySQL (www.mysql.org). Para a criação do modelo relacional foi utilizado o *software* livre DB Designer, que vem a ser uma ótima ferramenta criando as relações pertinentes à coesão do banco de dados. O MySQL é um Sistema Gerenciador de Banco de Dados (SGBD), que utiliza a linguagem SQL como interface, possui código aberto além de ser extremamente robusto e confiável.

3.3.2. Banco de Dados

No sentido de iniciar o processo de construção de um banco de dados proteogenômicos de *M. Tuberculosis* e devido à ausência de um banco de dados completo contendo todas as proteínas identificadas até o momento para o bacilo da TB, coletou-se os dados de estudos proteômicos independentes disponíveis na literatura em um banco de dados inicial de proteínas identificadas. Todos os dados proteômicos publicados anteriores ao início deste estudo, suportam a expressão de 1.586 proteínas, que representa um total de 40% de cobertura dos 3.947 genes codificadores de proteínas preditos e listados no banco de dados SwissProt (<http://www.expasy.ch/sprot/hamap/MYCTU.html>). Um único trabalho contribuiu com a maioria (1.044) destas proteínas, devido a utilização de uma abordagem baseada em LC-

MS/MS (MAWUENYEGA et al., 2005). Entretanto, devido à diversidade das abordagens utilizadas nestes estudos e falta de dados padronizados de identificação, os dados detalhados da literatura não puderam ainda ser adicionados ao banco de dados proprietário do presente projeto.

Após todo o processo de aquisição e identificação dos peptídeos e proteínas deste estudo, os dados obtidos foram estudados profundamente e, usando como modelo o esquema do banco de dados do sistema PEDRo, foi criada a estrutura relacional do sistema. As tabelas criadas, bem como as relações pertinentes às mesmas, estão representadas no esquema da Figura 7.

Até o presente momento, foram introduzidos no banco de dados um total de 143.411 identificações de alta qualidade provenientes da busca padrão dos dados. Estes dados serão constantemente atualizados a medida que os resultados das outras abordagens de busca sejam detalhadamente inspecionados.

A seguir, serão apresentadas as tabelas encontradas no modelo do banco de dados e uma apresentação das relações contidas entre elas.

Figura 9: Relação entre as tabelas do Banco de Dados Relacional do ProGen AP

3.4. Modelagem do Sistema

O estudo das necessidades e dos requisitos do sistema de banco de dados foi todo feito utilizando a linguagem UML (*Unified Modeling Language*). UML é uma linguagem de modelagem padronizada para a área de Engenharia de Software orientada a objeto. Essa linguagem proporciona uma forma-padrão para a preparação de planos de arquitetura de projetos incluindo, também, aspectos conceituais (BOOCH; JACOBSON; RUMBAUGH, 2006). Essa padronização foi criada e é gerenciada pela *Object Management Group* (OMG) e foi adicionada pela primeira vez à lista de tecnologias adotadas pela OMG em 1997. Desde então, a UML se tornou a ferramenta padrão para modelagem intensiva de sistemas.

No presente trabalho, a UML está relacionada aos procedimentos que devem ser executados pelos atores do sistema. A modelagem exemplifica a criação de novos usuários, a adição de novo conteúdo ao banco de dados e a utilização da ferramenta de busca de peptídeos, principais pontos do sistema.

3.4.1. Escopo do Sistema

O sistema ProGen AP é um sistema *Web* restrito apenas a usuários registrados e que oferece informações sobre grupos de peptídeos pertencentes a uma mesma proteína e que fazem parte do proteoma do organismo *Mycobacterium tuberculosis*. O acesso é controlado pela empresa VLS e apenas pessoas autorizadas possuem login/senha. As funcionalidades oferecidas pelo sistema, para cumprir os requisitos propostos, são descritas nas sessões subsequentes.

CADASTRAR USUÁRIOS

Novos usuários são cadastrados no sistema, pelo administrador. Todos os usuários estão armazenados em uma tabela separada no repositório de Banco de Dados do sistema. Para cadastrar um novo usuário, o administrador deve receber uma solicitação de cadastro via *email*. Essa requisição de cadastro chega ao administrador depois que o candidato a usuário do sistema preenche um formulário com informações básicas pessoais. Após essa notificação, o administrador acessa o banco de dados, em seguida acessa a tabela usuários e adiciona um novo usuário.

Existem apenas dois tipos de usuário nesse sistema: o usuário comum, com privilégios somente de acessar o banco de dados do *MTb*, e o usuário administrador, com privilégios de acesso ao banco de dados, criação de tabelas, inserção e remoção de dados. Desse modo, o

administrador pode designar, ainda, se esse usuário é usuário comum ou um administrador do sistema. Assim feito, a senha designada é enviada para o *email* do novo usuário cadastrado.

INSERIR CONTEÚDO

Uma função atribuída ao usuário Administrador é a possibilidade de inserir ou excluir conteúdo do banco de dados. A intenção é que a quantidade de informações a respeito do proteoma do organismo *MTb* aumente, de acordo com a quantidade de experimentos realizados, bem como com busca em outros sistemas repositórios já mencionados aqui anteriormente. Para inserir conteúdo, o Administrador precisa fazer *login* no banco de dados e, então, adicionar as novas linhas nas tabelas do banco.

É importante ressaltar aqui, que o usuário Administrador do sistema, também possui privilégios de administrador do Banco de Dados. Porém, para seguir as regras propostas na padronização MIAPE e não fugir do escopo desenvolvido baseado na arquitetura do PEDRo, recomenda-se que as tabelas propostas nesse sistema não devam ser alteradas, a menos em caso de um remodelamento do banco para comportar e integrar outros experimentos.

VISUALIZAR CONTEÚDO GERAL

Para visualizar os conteúdos do sistema é necessário possuir um usuário cadastrado. A visualização é em formato de tabelas, contendo uma lista de peptídeos, informações nominais sobre cada um deles (*score* e fração pertencente) e uma visualização geral (ferramenta de alinhamento gráfico) da localização desse peptídeo no genoma, bem como sobre qual fração aquele peptídeo pertence (de acordo com as cores de cada um - Extrato Total, Secretado, Condição Aeróbica ou Condição Anaeróbica).

VISUALIZAR CONTEÚDO ESPECÍFICO

Cada peptídeo possui informações específicas de carga, massa, sequência, posição no gene e no genoma, fração na qual foi identificado, entre outras. Para visualizar o conteúdo específico de um peptídeo é necessário possuir um usuário cadastrado, fazer a busca geral e depois, clicar no peptídeo de interesse.

3.4.2. Diagramas UML

Os diagramas são os meios utilizados para a visualização dos blocos de construção da UML, provêm uma parcial representação do sistema e ajudam a compreender a arquitetura

do sistema em desenvolvimento. Pode-se afirmar que o diagrama é uma representação gráfica de um conjunto de elementos que permitem a visualização do sistema a partir de diferentes perspectivas. Na UML existem vários tipos de diagramas utilizados na modelagem que especificam modelos a partir dos quais são construídos sistemas executáveis (BOOCH; JACOBSON; RUMBAUGH, 2006; PERSINOTI, 2009).

Diagrama de Casos de Uso

Define-se um diagrama de caso de uso como a representação gráfica mais genérica de uma interação de um usuário com o sistema. Tem a função de descrever a funcionalidade proposta para o sistema, através da descrição da sequência de eventos de um ator que usa o sistema para completar um determinado processo.

A Visão Global dos Atores e Casos de Uso é um diagrama que representa graficamente os atores do sistema e suas interações com os casos de uso. O ator, termo geralmente substituído por usuário, pode ser qualquer pessoa ou sistema ou qualquer outra forma de interação externa com o sistema em desenvolvimento. Caso de Uso é o conjunto de funcionalidades do sistema, representado por fluxos de eventos iniciados por um ator (CARDOSO, 2003).

Na Figura 10, pode-se observar o Diagrama de Caso de Uso: Visão Global dos Atores e Casos de Uso. Nesse diagrama são apresentados todos os atores e os casos de uso do sistema. O sistema possui apenas dois atores: Usuário comum e Administrador. Podem ser observados, no diagrama, os casos de uso do sistema, que são: Cadastrar Usuário, Inserir Dados no Banco, Busca no Banco e Pedir Acesso ao Banco. O diagrama mostra, ainda, as interações dos atores do sistema com cada um dos casos de uso, mostrando quais atores tem permissão para executar quais casos de uso.

Figura 10: Diagrama de Caso de Uso: Visão Geral dos Atores e Casos de Uso. Os atores Usuário Comum e Administrador são os Usuários do Sistema e interagem com os casos de uso que cada papel tem permissão para realizar. O ator visitante não é um usuário cadastrado no sistema, podendo visualizar somente os conteúdos públicos.

Diagramas de Colaboração

Um diagrama de colaboração exibe uma interação, consistindo de um conjunto de objetos e seus relacionamentos, incluindo as mensagens que podem ser trocadas entre eles. Esse tipo de diagrama dá ênfase à ordenação estrutural em que as mensagens são trocadas entre os objetos de um sistema.

A colaboração é representada por meio de diagramas que demonstram o relacionamento das classes no que diz respeito ao sentido das mensagens e ao número de mensagens, destacando a organização dos objetos que participam de uma interação (CARDOSO, 2003). Esse tipo de diagrama é útil pois permite observar o fluxo das informações e prever as possibilidades de ocorrência de problemas.

Na Figura 11, é possível observar o Diagrama de Colaboração **Pedir Cadastro**. Nesse diagrama são levantados todos os passos necessários para requerer um cadastro de novo usuário no sistema. Esse caso de uso é executado por um usuário comum, sem permissões de Administrador.

Figura 11: Diagrama de Colaboração do caso de uso Pedir Cadastro. Representa todos os eventos que o ator Usuário comum realiza para solicitar que tenha login e senha cadastrados no sistema.

O Diagrama de Colaboração **Cadastrar Usuário** é apresentado na Figura 12. Nesse diagrama são levantados todos os passos requeridos para cadastrar um novo usuário no sistema. É possível observar que apenas o Administrador tem permissão para realizar este caso de uso.

Figura 12: Diagrama de Colaboração do caso de uso Cadastrar Usuário. Representa todos os eventos que o ator Administrador realiza para criar um novo usuário do sistema.

O principal caso de uso do sistema está representado na Figura 13. Nesse caso de uso, denominado **Busca no Banco**, o ator Usuário (que pode ser usuário comum ou administrador) tem a possibilidade de efetuar buscas no banco de dados. Pode-se observar no diagrama o fluxo das informações e mensagens necessárias para efetuar uma busca com sucesso.

Figura 13: Diagrama de Colaboração do caso de uso Busca no Banco. Representa todos os eventos que o ator Usuário realiza para fazer uma busca por peptídeo ou proteína.

Além de efetuar buscas no banco, é necessário a possibilidade de poder inserir novos dados, garantindo que o repositório do sistema esteja sempre atualizado com dados de outros experimentos que irão agregar informações ao banco. Pensando nisso, existe o caso de uso representado na Figura 14, denominado **Inserir Dados no Banco**. Esse caso de uso só pode ser executado pelo ator Administrador. Além disso, o diagrama apresenta ainda os passos necessários para inserir novas informações no banco de dados.

Figura 14: Diagrama de Colaboração do caso de uso Inserir Dados no Banco. Representa todos os eventos que o ator Administrador realiza para poder inserir novos dados às tabelas do banco de dados do sistema.

**RESULTADOS E
DISCUSSÃO**

4. RESULTADOS E DISCUSSÃO

4.1. Análise dos Resultados

A coleta e a análise das amostras de MTb são parte de um projeto de iniciativa da empresa *Veritas Life Sciences*, cujo intuito principal foi identificar novos alvos para o diagnóstico precoce da tuberculose. O processo de identificação das amostras foi feito utilizando uma estratégia de fracionamento das amostras, visando a identificação de peptídeos pouco expressos ou expressos apenas em condições especiais. Dessa forma, fracionou-se a amostra em 5 sub-amostras, sendo elas: Extrato Total Anaeróbico, Extrato Total Aeróbico, Pellet, Secretado e Extrato total. Cada uma dessas sub-amostras, com exceção do Extrato Total, foram sub-divididas em 10 amostras, de acordo com sua posição no gel de eletroforese. Cada amostra recebeu uma nomenclatura diferente. A amostra de Extrato Total foi subdividida em 10 amostras, denominadas L01, L02 e assim sucessivamente até L10, e cada uma dessas sub-amostras também foi subdividida em 10 amostras. Resultando, então, em um total de 140 amostras, separadas por fração, de acordo com a Tabela 2.

Diversos algoritmos foram testados em todas as etapas do *pipeline* desse projeto. Todos deveriam obedecer alguns fatores, como: ser gratuito, pouca dificuldade de implementação, pouca ou nenhuma necessidade de manipulação no formato de entrada dos dados e, principalmente, a possibilidade de ser executado em uma máquina com pouco/médio poder computacional (pensando na portabilidade do mesmo). Assim sendo, alguns algoritmos testados durante a fase de experimentação, foram deixados de lado. Esse foi o caso dos algoritmos de busca *de novo* de identificação de peptídeos. Todos os algoritmos exigiram um tempo computacional muito grande (excederam 10 dias de processamento), além de muita manipulação de arquivos (mudança de mzXML para ms2 ou mgf, que não são os formatos padrões), geraram exceções de execução devido ao estouro de memória, além do fato de não terem sido muito eficientes para o conjunto de dados que foram analisados nesse projeto. Esse módulo pode ser integrado ao sistema, se o mesmo for executado em um servidor com configurações mais robustas e com maior poder de memória, porém o seu ganho em identificação de peptídeos ainda não foi confirmado.

Os dados tratados nesse projeto foram coletados através de um contrato de prestação de serviço em uma facility de análise proteômica no exterior, a fim de garantir a melhor qualidade analítica possível nesta etapa. O primeiro processamento foi feito pelo centro,

utilizando o LabKey, com o core X!Tandem, com os mesmos parâmetros da Busca Padrão, com um corte do PepProphet de 0.5 e com uma variação dMass menor que 20 partes por milhão. Visando então obter o máximo de identificações de peptídeos com *score* alto que também foram identificados no processamento inicial das amostras na facility e sabendo que o *software* LabKey disponível gratuitamente para *download* não possui os mesmos recursos que o do centro de processamento, efetuou-se um processamento inicial com uma fração das amostras, para definir os parâmetros de corte da análise local.

Partindo dessas informações, procedeu-se, então, duas abordagens: a primeira utilizando o parâmetro “*fractions*” na informação “*pipeline, data type*” (que analisa todos os arquivos mzXML juntos, como sendo de uma mesma amostra e faz um ranqueamento total dos dados) e a segunda, sem esse parâmetro e tratando cada uma das amostras independentemente, para verificar se o processamento com esse parâmetro agrega alguma informação a mais para a análise local. Os parâmetros utilizados nessa primeira etapa estão representados na Figura 15.

Tabela 2: relação das amostras por fração. Cada xx corresponde ao número de cada amostra, podendo variar de 01 ate 10.

Fração	Nomenclatura	Nº de Amostras
Extrato total anaeróbico	VLS_MTB_TeI_ANA_SLxx	10
Extrato total aeróbico	VLS_MTB_TeI_AER_SLxx	10
Pellet	VLS_MTB_PL_SLxx	10
Secretado	VLS_MTB_SE_SLxx	10
Extrato total**	VLS_MTB_TE_Lxx_SLxx	10 de cada

** cada L varia de 01 ate 10 e cada SL, varia de 01 ate 10.


```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <bioml>
3   <note label="residue, potential modification mass" type="input">242.101505@K,3.0188302@C,15.994915@M</note>
4   <note label="residue, modification mass" type="input">71.037114@C</note>
5   <note label="protein, cleavage site" type="input">[KR]||P</note>
6   <note label="spectrum, parent monoisotopic mass error plus" type="input">2.0</note>
7   <note label="spectrum, parent monoisotopic mass error minus" type="input">2.0</note>
8   <note label="pipeline, data type" type="input">fractions</note>
9   <note label="scoring, maximum missed cleavage sites" type="input">4</note>
10  <note label="scoring, algorithm" type="input">k-score</note>
11  <note label="spectrum, use conditioning" type="input">no</note>
12  <note label="scoring, minimum ion count" type="input">1</note>
13  <note label="pipeline, protocol name" type="input">BuscaPadrao_fractions</note>
14  <note label="pipeline, protocol description" type="input"/>
15  <note label="pipeline, database" type="input">Mycobacterium_tuberculosis_H37Rv.fasta</note>
16 </bioml>
17

```

Figura 15: parâmetros do X!Tandem incluindo “fractions” na análise preliminar dos dados.

Para conseguir obter o máximo possível de identificações de peptídeos com *score* alto (maior que 0.7) e que haviam sido identificados anteriormente, pela *facility*, foi utilizada uma fração do total de amostras para fazer um processamento inicial, controlando o tempo de análise de cada uma das amostras, bem como o resultado de cada uma delas. Utilizaram-se as frações VLS_MTB_TE_L8_SL01 até VLS_MTB_TE_L8_SL10 (equivalente a aproximadamente 8% do total de amostras).

Tabela 3: Quantidade de peptídeos identificados na análise com o parâmetro “fractions” e sem, com as amostras sendo analisadas individualmente.

	01_to_10	fractions
Pept. identificados	31.758	31.532
Pept. Iguais a Seattle	29.786	29.536
% com pepProphet: 0,5<x<0,6	54,04%	55,13%
% com pepProphet: 0,6<x<0,7	68,58%	71,44%
% com pepProphet: 0,7<x<0,8	79,90%	77,80%
% com pepProphet: 0,8<x<0,9	88,18%	88,10%
% com pepProphet: 0,9<x<1,0	99,26%	99,25%

A partir do resultado da Tabela 3, decidiu-se, então, somente considerar espectros identificados que possuem um *score* dado pelo PepProphet acima de 0.7, o que garante uma

taxa de erro menor que 5% (0.0425), considerada aceitável pelo pesquisador especialista. A curva de Sensitividade e Erro pode ser vista na Figura 16.

Figura 16: curva de sensibilidade e taxa de erro da análise preliminar dos dados

4.2. Identificação dos peptídeos

O processo de identificação de peptídeos nos dados de espectrometria de massa iniciou-se pelo método padrão de identificação, onde os espectros são correlacionados com espectros simulados para peptídeos trípticos presentes em um banco de dados de proteína. Esta análise foi realizada utilizando o algoritmo X!Tandem (algoritmo “open source” incorporado no sistema LabKey (NELSON et al., 2011)) e o banco de dados de proteínas preditas no genoma da cepa H37Rv do *MTb*, disponível no GeneBank (VERSION NC_000962.2 de 09/Nov/2009). Apesar de poderem gerar mais identificações, não foram considerados, nesta primeira rodada de análises, os peptídeos semi-trípticos, que apresentam sítio para a tripsina em apenas uma das extremidades, uma vez que o tempo de processamento destes resultados é proporcionalmente maior de acordo com a combinação de possibilidades de peptídeos com apenas uma ligação tríptica em uma das extremidades.

Desta primeira etapa da análise, identificou-se 144.146 espectros, correspondendo a 129.485 peptídeos, de acordo com a Tabela 4. Existe essa diferença, pois, como explicado anteriormente, um mesmo peptídeo pode estar representado por mais de um espectro, gerando uma certa redundância nos resultados obtidos. Isso gera uma necessidade de filtragem dos espectros, para encontrar o número de peptídeos únicos. Para tal efeito, decidimos utilizar

como peptídeo único, um peptídeo que tenha a fórmula “scan_sequência” aparecendo apenas uma única vez na lista de peptídeos identificados. Peptídeos redundantes foram desconsiderados. A partir dessa lista de peptídeos únicos, conseguiu-se identificar um total de 2.227 proteínas. Todos esses peptídeos foram adicionados ao banco de dados do ProGen AP.

Tabela 4: Relação de quantidade de identificações resultantes na Busca Padrão.

	# identificações
Espectros	144.146
Peptídeos	129.485
Proteínas	2.227

A estratégia seguinte abordada foi a Busca Refinada, na qual buscou-se também por peptídeos que fossem semi-trípticos, (ou seja, apenas uma das extremidades previsa como sendo produto de clivagem pela tripsina e a outra sendo uma clivagem inespecífica), ampliando o campo de busca. Dessa busca obtiveram-se 133.253 peptídeos totais, sendo que 13.435 desses eram exclusivos da Busca Refinada e foram acrescentados, também, à lista de peptídeos presentes no sistema repositório de banco de dados deste projeto. O resultado dessa etapa pode ser visto na Tabela 5.

Tabela 5: Relação de quantidade de identificações resultantes na BuscaRefinada

	# identificações
Espectros	147.922
Peptídeos	133.253
Pep. acrescentados	+13.435
Proteínas acresc.	+82

A Busca Genômica também resultou em identificações novas (conforme representado na Tabela 6), acrescentando ainda 13.838 novas identificações de peptídeos à lista total. É importante ressaltar que o processo de identificação da Busca Genômica é um pouco mais pesado em termos de processamento computacional e estratégia, pois necessita de uma tabela

com a localização daquele fragmento identificado no genoma, portanto o processamento consiste em, além de fazer a comparação da sequência com o espectro, fazer a relação daquela sequência com a localização no genoma.

Tabela 6: Relação de quantidade de identificações resultantes na Busca Genômica

	# identificações
Espectros	117.569
Peptídeos	117.569
Pep. acrescidos	+13.838

A busca em Bibliotecas de Espectros gerou, em sua maioria, identificações que, apesar de corresponderem com a biblioteca criada, não obtiveram *scores* considerados confiáveis. Propôs-se então, de posse dos resultados oriundos do QuickMod, que os espectros sequenciados passassem pela análise do PeptideProphet, para que, o algoritmo criasse sua própria curva de aprendizado, padronizando os *scores* de todas as etapas e possibilitando uma comparação de medidas de identificação de peptídeos.

O resultado final foi a identificação de 232 espectros que não haviam sido identificados antes e que, com o processamento feito com o QuickMod seguido do PeptideProphet, foram identificados com o padrão de confiabilidade anterior (pepprophet $\geq 0,7$).

Figura 17: resultado geral de identificação de peptídeos

Dentre todos os resultados aqui apresentados e representados na Figura 17, é importante ressaltar que estes dados obtidos apresentam grau de confiança superior a 95%, com sensibilidade média de 0,7725 e taxa de erro de 0,0473. O alto grau de confiança aqui alcançado é devido à utilização de critérios refinados para identificação, como a análise estatística obtida através da ferramenta PeptideProphet (KELLER et al., 2002), parte integrante do sistema LabKey. Todos estes peptídeos foram inseridos no banco de dados e já estão disponíveis para consulta.

A primeira observação importante é que dentre todas as proteínas até o momento, aqui identificadas, a maioria delas está anotada como proteínas hipotéticas (“*putative*”, “*probable*” ou “*hypothetical*”). Esses dados suportam a existência destes produtos gênicos. Além disso, identificaram-se 567 peptídeos N-terminal (26 em proteínas hipotéticas) e 1.229 peptídeos Cterminal (319 em proteínas hipotéticas), indicando a correta predição dos pontos de início e interrupção da tradução deste genes. Estas informações podem ser diretamente introduzidas em bancos de dados como o GeneBank para completar a anotação do genoma.

4.3. O sistema

A utilização das ferramentas adequadas e a análise dos objetivos desse projeto possibilitaram a construção de uma interface amigável, intuitiva e, ao mesmo tempo, robusta para a pesquisa e visualização dos dados do sistema. Ilustram-se nas figuras a seguir cada parte da interface do sistema e seu funcionamento, baseado na sequência de acesso e utilização do mesmo (Inicial, Pesquisa, Resultado Geral e Resultado Específico). Por se tratar de um sistema com interface *web*, a forma como foram disponibilizadas as informações foi feita de maneira a deixar um aspecto limpo e intuitivo. Todas as imagens (captura de tela) a seguir foram obtidas do sistema funcional disponível no endereço www.veritasls.com.br/mtbproteome.html.

Pôde-se perceber com a utilização do sistema para testes, que o mesmo se tornou ágil e de leve visualização, apesar da quantidade de informações que o mesmo processa em cada consulta. A programação PHP e os *scripts* SQL tornaram o sistema seguro ao acesso e à disponibilização das informações nele contidas.

Algumas decisões de projeto, como por exemplo, restringir a inserção de novos dados ao curador do banco (pesquisador responsável), asseguram a qualidade das informações nele disponibilizada. Desse modo, apenas dados cuja proveniência, confirmada pelo pesquisador, seja idônea serão inseridos no repositório. O refinamento das tabelas e a criação de tabelas

referentes a informações específicas de cada experimento, ampliando o banco para um repositório geral de proteomas de diferentes organismos, são, também, metas futuras para esse sistema.

4.3.1. Página Inicial

Ao acessar o sistema, a primeira página avistada por todos é uma página com um breve resumo a respeito do projeto (Figura18) e logo abaixo um resumo dos resultados obtidos, bem como informações sobre o que está disponível para ser acessado (Figura 19).

Figura 18: Página com Introdução e Motivação do projeto

Figura 19: Continuação da Página de Introdução, agora com resumo do projeto

Essa área inicial pode ser acessada por qualquer pessoa que esteja apenas visitando. Nessa área, a pessoa visitante pode obter informações sobre a abordagem experimental utilizada nesse estudo, quantidade de espectros obtidos e quantidade de peptídeos identificados.

4.3.2. Área de Acesso

Quando na Página Inicial, clica-se no link “*Click here to search in the database*”, o visitante é direcionado para a Área de Acesso. Essa Área de Acesso está representada na Figura 20. No canto inferior da tela existe um aviso de que esse sistema é um sistema fechado e com acesso restrito, fornecendo um *link* logo em seguida para que o usuário possa solicitar o acesso ao sistema.

MTB Proteome

ÁREA DE ACESSO

Login :

Senha:

ATENÇÃO: Este é um Banco de Dados particular e com acesso restrito e controlado. Se você não é um usuário cadastrado [CLIQUE AQUI](#).

Figura 20: Página de Login/Senha

4.3.3. Pesquisa no Banco de Dados

Depois que o usuário é cadastrado, ele pode fazer o seu *login* na Área de Acesso e, assim, ser redirecionado para a área de Pesquisa no Banco de Dados (Figura 21). Essa área possui um campo para o usuário inserir a sua palavra-chave (*QUERY*), podendo ser tanto a sigla do gene (*gene symbol*), a etiqueta de localização (*locus tag*) ou a descrição do gene (*gene description*).

Todas as buscas devem seguir padrões de nomenclatura do NCBI (Entrez Gene) e, caso o usuário tenha alguma dúvida sobre “como procurar”, no canto inferior da tela existe um exemplo e o link para o site do Entrez Gene, contendo as informações adequadas.

MTB Proteome

SEARCH IN DATABASE

Query:

- Select -

- Select -

Gene Symbol

Locus Tag

Gene Description

HOW TO SEARCH: all the tags used in the query construction is based in the standards of the [Entrez Gene](#) Summary Result, from [NCBI](#). Click [here](#) to see an example (*Gene Symbol: dnaA - Locus Tag: Rv0001*).

Discover • Innovate • Transform

Figura 21: Página principal de Busca

Assim que o usuário inserir uma palavra-chave de interesse e clicar em “Search” na Área de Busca, deverá aparecer uma tela com a lista de resultados. Nesse caso, o usuário fez a busca pela sigla do gene *dnaN*. A página de Resultado Geral está representada na Figura 22 e foi dividida em três partes:

- A) A primeira parte contém as informações gerais sobre o gene procurado (*Gene Name*, *Locus Tag*, *Acess Number*, *Description*). Essas são informações obtidas do banco de dados genômico.
- B) Na segunda parte é possível visualizar informações do banco de dados proteômico. Nessa área são listados todos os peptídeos cuja localização esteja relacionada ao produto proteico daquele gene em questão. As informações foram colocadas em formato de tabela, podendo ser ordenada de acordo com o critério de maior *score* ou pelo tipo de fração na qual o peptídeo foi encontrado.
- C) Na terceira parte dessa página é possível ver a ferramenta de visualização mostrando a posição do gene no genoma e a posição de cada peptídeo dentro daquele gene.

4.3.5. Resultado Específico

Após obter informações a respeito de todos os peptídeos identificados relacionados ao gene de interesse, o usuário pode ainda obter informações específicas sobre um determinado peptídeo. Essa informação é útil quando se compara a expressão proteica em duas condições diferentes (aeróbica e anaeróbica, por exemplo), pois existem peptídeos que são expressos somente em uma das condições, podendo vir a ser um potencial marcador biológico. No contexto biológico desse estudo, identificar peptídeos com alta taxa de confiabilidade (*score* >0.7) expressos somente em condições anaeróbicas e não em condições aeróbicas nos permite fazer uma lista de possíveis marcadores para *MTb* em estado latente (fase anaeróbica do organismo).

Essa página está representada na Figura 23 e também é dividida em três partes, sendo:

- A) Informações de quantificação do peptídeo identificado;
- B) Sequência do peptídeo identificado;
- C) Ferramenta Gráfica de Visualização do Alinhamento.

Figura 23: Página Visualização do Resultado de Busca por Peptídeo Específico

Um componente importante do sistema é a ferramenta de alinhamento e visualização gráfica dos dados. Este componente combina todas as informações listadas na tabela e garante toda a intuitividade do alinhamento de sequências para a análise detalhada dos dados. Este módulo gráfico para visualização do alinhamento foi construído nos moldes da ferramenta apresentada no PeptideAtlas - <http://www.peptideatlas.org/software.php>, utilizando linguagem Perl e bibliotecas gráficas de BioPerl.

Um exemplo do tipo de alinhamento que pode ser visualizado em uma busca por um determinado gene está ilustrado na figura 24.

Figura 24: Exemplo da ferramenta de visualização gráfica de alinhamentos

5. CONCLUSÕES

Sistemas computacionais devidamente projetados para área biotecnológica tem se mostrado cada vez mais como uma importante ferramenta de auxílio à extração de novos conhecimentos a respeito da análise proteômica vários organismos-alvo, com grandes perspectivas de aprimoramento e refinamento de seus métodos. O desenvolvimento desse sistema vem de encontro a uma nova necessidade em se tratando de pesquisas biológicas: a criação de sistemas de banco de dados particularizados para algumas áreas (a proteogenômica dentre elas), subdividindo e filtrando as grandes bases de dados e adaptando-as às necessidades de grupos de pesquisa em áreas específicas, porém com grande capacidade de integração e de troca de informações com outros sistemas. Além disso, sistemas *Web* sevem perfeitamente a esse propósito, pois são muito abrangentes, sendo independentes da plataforma utilizada pelo usuário final e facilmente acessado em qualquer lugar, bastando apenas ter acesso a *Internet*.

A construção de um repositório de dados bem planejado, com arquitetura sólida e bem consistente baseada em repositórios já reconhecidamente eficientes, como o PEDRo, possibilita a realização de operações de “*data mining*” e a obtenção de novos conhecimentos com alto nível de qualidade e confiança. O sistema ProGen AP foi desenvolvido utilizando os padrões requisitados e está disponível para o acesso de pesquisadores e colaboradores do projeto (mediante solicitação e aprovação de acesso pela *Veritas Life Sciences*) através do endereço <http://www.veritasls.com.br/mtbproteome.html>. O sistema oferece informações a respeito de experimentos proteômicos com o organismo *Mycobacterium tuberculosis*.

O sistema foi testado pelos pesquisadores-especialistas e funcionários colaboradores da empresa *Veritas Life Sciences*. Após os testes, as mudanças sugeridas foram implementadas e então foi liberado o cadastro de usuários externos, mediante autorização. Por fazer parte de um projeto temático maior, desenvolvido pela empresa VLS, sob supervisão da Dr^a Sandra R.Pereira Faça, o sistema foi também acessado e testado pelos acesores responsáveis pela avaliação do projeto FAPESP.

Pode-se afirmar que o estudo e o desenvolvimento do *pipeline* para anotação de peptídeos e proteínas do organismo estudado teve elevado sucesso, tendo identificado aproximadamente 57% de todas as proteínas do organismo. Essa taxa de cobertura das identificações do *pipeline* proposto já se equivale e supera resultados obtidos em aparelhos de espectrometria de massa de última geração, nos quais a taxa de identificação é aproximadamente 45%.

O sistema ProGen AP, assim como qualquer sistema, poderá sofrer alterações decorrentes de atualizações e manutenções que se façam necessárias, visando o crescimento e possivelmente, a incorporação de outras funcionalidades e outros tipos de dados, para complementá-lo. Uma das alterações a ser feita, será a migração para um servidor robusto, possibilitando o aumento do volume de dados. Outra alteração, será, após a migração, a integração de um módulo de busca *de novo*, para analisar os espectros restantes e tentar identificar mais proteínas, aumentando a cobertura desse experimento.

O sistema ProGen AP, em um primeiro momento, não será disponibilizado ao público, porém espera-se que o sistema e as informações geradas nesse projeto sejam de grande utilidade aos pesquisadores envolvidos no projeto, ajudando a descoberta de potenciais alvos de medicamentos contra a Tuberculose, bem como diagnósticos mais precoces.

6. REFERÊNCIAS BIBLIOGRÁFICAS

- AHRNÉ, E.; MULLER, M.; LISACEK, K. **Unrestricted identification of modified proteins using MS/MS**. *Proteomics*, ed.10, p.671-686, 2010.
- AHRNÉ, E.; NIKITIN, F.; LISACEK, F.; MULLER, M. **QuickMod – a tool for open modification spectrum library searches**. *Journal Proteome Research*, ed.10, p.2913-2921, 2011.
- BEAUSOLEIL, S.A. et al. **Large-scale characterization of HeLa cell nuclear phosphoproteins**. *Proc.Natl. Acad. Sci. Ed.*101, p.12130-12135, 2004.
- BOOCH, G.; JACOBSON, I.; RUMBAUGH, J. **UML, Guia Prático do Usuário**. Tradução. Rio de Janeiro. Campus. 2006.
- BRAZMA, A. et al. **Minimum information about a microarray experiment (MIAME) - toward standards for microarray data**. *Nat Genet*, ed.29, p.365–371, 2001.
- CARDOSO, C. **UML na prática: do problema ao sistema**. Rio de Janeiro. Editora Ciência Moderna Ltda. 2003.
- CONVERSE, T.; PARK, J. **PHP4, a Bíblia**. Ed. Campus, 2ª edição, 2000.
- COX, J.; MANN, M. **Is proteomics the new genomics?** *Cell*, ed 130, p.395-398, 2007.
- CRAIG, R.; BEAVIS, R.C. **A method for reducing the time required to match protein sequences with tandem mass spectra**. *Rapid Commun. Mass Spectrom.* ed.17(20), p.2310-2316, 2003.
- CRAIG, R.; BEAVIS, R.C. **TANDEM: matching proteins with tandem mass spectra**. *Bioinformatics.*, Jun 12, ed.20(9), p.1466-7, 2004.
- CRAVATT, B.F.; SIMON, G.M.; YATES, J.R. **The biological impact of mass-spectrometry-based proteomics**. *Nature*, 3rd, ed 450, p.991-1000, 2007.
- DEUTSCH, E.W. et. al. **A guided tour of the TransProteomic Pipeline**. *Proteomics*, ed.10, vol.6, p.1150, 2010.
- ENG, J.K.; McCORMACK, A.L.; YATES, J.R. **An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database**. *J Am Soc Mass Spectrom*, ed.5 (11), p.976–989, 1994.
- FACA, V. et al. **Contribution of protein fractionation to depth of analysis of the serum and plasma proteomes**. *J Proteome Res*, ed 6, p.3558-3565, 2007.
- FALKNER, J.A. et al. **A Spectral Clustering Approach to MS/MS Identification of Post-Translational Modifications**. *Journal of Proteome Research*, 2008, ed 7, p.4614-4622.

FENIO, D.; BEAVIS, R.C., **A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes.** Anal. Chem. Feb 15, ed.75(4), p.768-774, 2003.

FICARRO, S.B. et al. **Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*.** Nat. Biotechnol., ed.20, p.301-305, 2002.

FRANK, A.M. et al. **De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry.** Journal of Proteome Research, ed.6, p.114-123, 2007.

FRANK, A.M. et al. **Clustering Millions of Tandem Mass Spectra.** Journal of Proteome Research, ed.7, p.113-122, 2008.

FREWEN, B; MACCOSS, M.J. **Using BiblioSpec for Creating and Searching Tandem MS Peptide Libraries.** Curr. Protoc. in Bioinform., 20, p. 13.7.1-13.7.12, 2007.

GEER, L.Y. et al. **Open Mass Spectrometry Search Algorithm.** J. Proteome Res. Ed.3(5), p.958-964, 2004.

GUELICH, S.; GUNDAVAN, S. **Programação CGI com Perl** (O Reilly). Rio de Janeiro, Editora Ciência Moderna Ltda 2001.

GARWOOD, K.L. et al. **PEDRo: A Configurable Data Entry Tool for XML.** Bioinformatics Oxford University Press, vol 20, ed. 15, p.2463-2465, 2004.

GUPTA, N. et al. **Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation.** Genome Res, ed 17, p. 1362-1377, 2007.

HAN, D.K et al. **Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry.** Nature Biotechnology, ed.19(10), p.946-951.2001,

HANASH S.; PITTERI, S; FACA, V. **Mining the plasma proteome for cancer biomarkers.** Nature, Apr 3; ed. 452(7187), p. 571-579, 2008.

HOOGLAND, C. et al. **SWISS-2DPAGE, ten years later.** Proteomics, ed.4(8), p.2352-2356, 2004.

HUNT, D.F. et al. **Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry.** Science. Mar 6; ed.255(5049), p.1261-1263, 1992.

JAFFE, J.D.; BERG, H.C.; CHURCH G.M. **Proteogenomic mapping as a complementary method to perform genome annotation.** Proteomics, ed 4, p.59-77, 2004.

KÄLL, L., et al. **The optimal linear combination of the list of scoring features is derived using a Support Vector Machine classification algorithm.** Nature Methods, ed.4, p.923-925, 2007.

KAUFMANN, S.H., **How can immunology contribute to the control of tuberculosis?** Nat Rev Immunol ed.1, p. 20-30, 2001.

KAUFMANN, S.H. **Robert Koch, the Nobel Prize, and the ongoing threat of tuberculosis.** N Engl J Med ed.353, p.2423-2426, 2005.

KELLER, A. et al. **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.** Anal Chem ed.74, p.5383-5392, 2002.

MANN, M.; JENSEN, O. **Proteomic analysis of post-translational modifications.** Nature Biotechnology, ed.21, p.255-261, 2003.

MAWUENYEGA, K.G. et al. **Mycobacterium tuberculosis functional network analysis by global subcellular protein profiling.** Mol Biol Cell ed.16, p. 396-404, 2005.

MOLLENKOPF, H.J. et al. **A dynamic twodimensional polyacrylamide gel electrophoresis database: the mycobacterial proteome via Internet.** Electrophoresis ed.20, p.2172-2180, 1999.

NELSON, E.K. et al. **LabKey Server: An open source platform for scientific data integration, analysis and collaboration.** BMC Bioinformatics, ed.12, p.71, 2011.

NESVIZHSKII, A.I. et al. **A statistical model for identifying proteins by tandem mass spectrometry.** Anal Chem ed.75, p.4646-4658, 2003.

OLIVEIRA, C.H.P. **SQL - Curso Prático**, 2002, Novatec Editora.

PARKINSON, H. et al. **ArrayExpress—a public database of microarray experiments and gene expression profiles.** Nucleic Acids Res, ed. 35, p. D747–D750, 2007

PANDEY, A. **Proteomics to study genes and genomes.** Nature, ed 405, p.837-846, 2000.

PERKINS, D.N. **Probability-based protein identification by searching sequence databases using mass spectrometry data.** Electrophoresis, vol 20, ed. 18, p.3551-3567, 1999.

PERSINOTI, G.F. **Auxílio ao Diagnóstico de Artrite Reumatóide Através de Técnicas de Inteligência Artificial.** Dissertação de Mestrado, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo. Ribeirão Preto. 2009.

QUACKENBUSH, J. **Standardizing the standards.** Mol Syst Biol 2006, 2: 2006–2010.

RAUCH, A. et al. **Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments.** J Proteome, Res 5, p.112-121, 2006.

ROMBEL, I.T. et al. **ORF-FINDER: a vector for high-throughput gene identification.** Gene, vol. 282, p. 33-41, 2002.

SEARLE, B.C., **An explanation of the Peptide Prophet algorithm developed by Keller,A., Nesvizhskii, A.I., Kolter, E., and Aebersold, R..** www.proteomesoftware.com.

STEEN, H. et al. **Tyrosine phosphorylation mapping of the epidermal growth factor receptor signaling pathway.** J.Biol.Chem., ed.277, p.1031-1039, 2002.

STEIN, L. **How Perl saved the human genome project.** The Perl Journal 1, 1996. http://www.bioperl.org/wiki/How_Perl_saved_human_genome. Retrieved 2009-02-25 (Bioperl.org). BioPerl: Main Page – Disponível em <http://www.bioperl.org/> .

STEIN, E.S.E.; RUDNICK, P.A. **NIST Peptide Tandem Mass Spectral Libraries.** E.coli peptide Mass Spectral Reference Data, *E.coli*, ion trap, qtof, data oficial de construção> 4 de fevereiro de 2009. National Institute of Standards and Technology, Gaithersburg, MD, 20899. Download feito em 29 de julho de 2012, do site <http://peptide.nist.gov>.

TAYLOR, C.F. et al. **A systematic approach to modeling, capturing, and disseminating proteomics experimental data.** Nat Biotechnol, ed 21, p.247–254, 2003.

TAYLOR, C.F. et al. **The minimum information about a proteomics experiment (MIAPE).** Nat Biotechnol, ed 25, p.887–893, 2007.

TISSOT, B. et al. **Glyco-Proteomics: past, present and future.** FEBS Lett., 583, p.1728-1735, 2009.

ZHANG, J. et al. **PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification.** Mol Cell Proteomics, ed.11:10.1074/mcpM111.010587, p.1-8, 2012.

Get Started with MySQL - Disponível em http://dev.mysql.com/usingmysql/get_started.html .

WHO (2006) World Health Organization - Global tuberculosis control - surveillance, planning, financing. http://www.who.int/tb/publications/global_report/2006/en/.

Programa	www	Disponibilidade
Programa de Busca em Bancos de Dados		
SEQUEST	thermo.com	-
MASCOT	matrixscience.com	*
ProteinProspector	prospector.ucsf.edu	*
ProbiD	tools.proteomecenter.org/wiki/index.php?title=Software:ProbiD	**
X!Tandem	www.thegpm.org	*, **
SpectrumMill	www.chem.agilent.com	-
Phoenyx	www.genebio.com/products/phenyx	-
OMSSA	pubchem.ncbi.nlm.nih.gov/omssa	*, **
VEMS3.0	yass.sdu.dk	**
ProteinPilot	www.absciex.com	-
MyriMatch	fenchurch.mc.vanderbilt.edu/software.php	**
PepSlice	www.ti.inf.ethz.ch/pw/software/pepslice	**
RAId_DbS	www.ncbi.nlm.nih.gov/CBBresearch/qmbp/raid_dbs/	*, **
MassMatrix	www.massmatrix.net/mm-cgi/home.py	-
Programas de Spectral Matching/Busca em Bibliotecas de Espectros (<i>Spectral Library</i>)		
SpectraST	www.peptideatlas.org/spectrast/	*, **
X!P3	p3.thegpm.org/tandem/ppp.html	*
QuickMod	http://javaproplib.sourceforge.net/packages/tools/quickmod/index.html	*, **
Bibliospec	proteome.gs.washington.edu/software/bibliospec/	-
Ferramentas de Bancos de dados de MS/MS		
PeptideAtlas	www.peptideatlas.org	
Proteios	www.proteios.org	
SBEAMS	sbeams.org	
CPAS	www.labkey.org	
PRIDE	www.ebi.ac.uk/pride/	
Peptidome	www.ncbi.nlm.nih.gov/peptidome/	
MASPECTRAS2	genome.tugraz.at/maspretras	
Ferramentas com Abordagens Híbridas		
InsPecT	proteomics.ucsd.edu/Software/Inspect.html	*, **
Popitam	www.expasy.org/tools/popitam/	**
TagRecon	fenchurch.mc.vanderbilt.edu/software.php	**
ByOnic	www.parc.com/work/focus-area/mass-spectra-analysis/	-
SpectralNetworks	proteomics.ucsd.edu/Software/SpectralNetworks.html	**
MODi	www.massmatrix.net/mm-cgi/home.py	-

*acesso gratuito via internet

** distribuição gratuita

h:

ANEXO B - Exemplo do arquivo do Banco de Dados Genômico

```
>MycoTub_F1_00001
LTDDPGSGFTTVWNAVVSSELNGDPKVDDGPSSDANLSAPLTPQQRAWLNL
VQPLTIVEGFALLSVPSSFVQNEIERHLRAPITDALSRRLGHQIQLGVRI
APPATDEADDTTVPPSENPAATTSPDTTTDNDEIDDSAAARGDNQHSWPSY
FTERPHNTDSATAGVTSLNRRYTFDTFVIGASNRFAHAAALAI AEAPARA
>MycoTub_F1_00002
APPATDEADDTTVPPSENPAATTSPDTTTDNDEIDDSAAARGDNQHSWPSY
FTERPHNTDSATAGVTSLNRRYTFDTFVIGASNRFAHAAALAI AEAPARA
YNPLFIWGESGLGKTHLLHAAGNYAQRLLFPGMRVKYVSTEEFTNDFINSL
RDDRKVAFKRSYRDVDVLLVDDIQFIEGKEGIQEEFFHTFNTLHNANKQI
>MycoTub_F1_00003
YNPLFIWGESGLGKTHLLHAAGNYAQRLLFPGMRVKYVSTEEFTNDFINSL
RDDRKVAFKRSYRDVDVLLVDDIQFIEGKEGIQEEFFHTFNTLHNANKQI
VISSDRPPKQLATLEDRLRTRFEWGLITDVQPPELETRIAILRKKQMER
LAVPDDVLELIASSIERNIRELEGALIRVTAFASLNKTPIDKALAEIVLR
>MycoTub_F1_00004
VISSDRPPKQLATLEDRLRTRFEWGLITDVQPPELETRIAILRKKQMER
LAVPDDVLELIASSIERNIRELEGALIRVTAFASLNKTPIDKALAEIVLR
DLIADANTMQISAATIMAATAEYFDTTVEELRGP GKTRALAQSRQIAMYL
CRELTDLSLPKIQGAQFGRDHTTVMYAQRKILSEMAERREVFHDHVKELTTR
>MycoTub_F1_00005
DLIADANTMQISAATIMAATAEYFDTTVEELRGP GKTRALAQSRQIAMYL
CRELTDLSLPKIQGAQFGRDHTTVMYAQRKILSEMAERREVFHDHVKELTTR
IRQSKRHGVFFRQSKNFSLPGHTSHRDWLVS LCTNRAQTHTVPAVPFT
THASSPPTQHTPHSHRHRHPQLRPTSTCTKTRLS PNCTPSNTVTEISSSF
>MycoTub_F1_00006
IRQSKRHGVFFRQSKNFSLPGHTSHRDWLVS LCTNRAQTHTVPAVPFT
THASSPPTQHTPHSHRHRHPQLRPTSTCTKTRLS PNCTPSNTVTEISSSF
VLGKTALGIVRWIPPALARRGGSEVNDELSSREKLYGCCSTAVAAVVAGH
ASWAFVVGSPHASGALATGSNSSGERAQWTRLRQELASPTRFVCYESLSP
>MycoTub_F1_00007
VLGKTALGIVRWIPPALARRGGSEVNDELSSREKLYGCCSTAVAAVVAGH
ASWAFVVGSPHASGALATGSNSSGERAQWTRLRQELASPTRFVCYESLSP
MRCRGWLKICQPGPRCRCSPACCPARTTVRFDPDSTTRFP RPRLALKLFL
LEAFFLADCCPILPGRCLTSPTFMSKVTGSHPAVTPGFRYRRCQSRIIRR
>MycoTub_F1_00008
MRCRGWLKICQPGPRCRCSPACCPARTTVRFDPDSTTRFP RPRLALKLFL
LEAFFLADCCPILPGRCLTSPTFMSKVTGSHPAVTPGFRYRRCQSRIIRR
CRRRCRKPDCCLRNYSRQSVRSLSPPAGTTRCLCPASGSKSSVRRWFWP
LPTGFAWLFENSRRRRQISKRLCWSRPRRWPRPPKRASAALTFVCRWVL
>MycoTub_F1_00009
CRRRCRKPDCCLRNYSRQSVRSLSPPAGTTRCLCPASGSKSSVRRWFWP
LPTGFAWLFENSRRRRQISKRLCWSRPRRWPRPPKRASAALTFVCRWVL
GRGWARMACSVSVGTASAAPRDFLMPSSRSFGSCYQPNTPRWPPWTWPSS
KRSSWLRLWIGARRCAWSSLMAACGFLRVPMMLDEPRKILLLTMPVNHRL
>MycoTub_F1_00010
GRGWARMACSVSVGTASAAPRDFLMPSSRSFGSCYQPNTPRWPPWTWPSS
KRSSWLRLWIGARRCAWSSLMAACGFLRVPMMLDEPRKILLLTMPVNHRL
RLTQPIRTVVR CARSECLSGRLRLVSLPCYVRCPGTIALWRVMATVRSRR
CRRTMSICCRFGCRAEHLAPGVYVRHLGLRDFRSWACVDLELHPGRTV FV
GPNYGKTNLIEALWYSTTLGSHRVSADLPLIRVGTDRAVISTIVVNDGR
ECAVDLEIATGRVNKARLNRS SVRSTRDVVGVLRAVL FAPEDLGLVRGDP
>MycoTub_F1_00012
GPNYGKTNLIEALWYSTTLGSHRVSADLPLIRVGTDRAVISTIVVNDGR
ECAVDLEIATGRVNKARLNRS SVRSTRDVVGVLRAVL FAPEDLGLVRGDP
ADRRRYLDDLAIVRRPAIAAVRAEYERVLRQRTALLKSVPGARYRGDRGV
```


FDTLEVWDSRLAEHGAELVAARIDLNVQLAPEVKKAYQLLAPESRSASIG
 >MycoTub_F1_00013
 ADRRRYLDDLAIVRRPAIAAVRAEYERVLRQRTALLKSVPGARYRGDRGV
 FDTLEVWDSRLAEHGAELVAARIDLNVQLAPEVKKAYQLLAPESRSASIG
 YRASMDVTGTPSEQSDIDRQLLAARLLAALAARRDAELERGVCLVGPHRDD
 LILRLGDQPAKGFASHGEAWSLAVALRLAAYQLLRVDGGEPVLLLDVFA
 >MycoTub_F1_00014
 YRASMDVTGTPSEQSDIDRQLLAARLLAALAARRDAELERGVCLVGPHRDD
 LILRLGDQPAKGFASHGEAWSLAVALRLAAYQLLRVDGGEPVLLLDVFA
 ELDVMRRRALATAAESAEQVLVTAAVLEDIPAGWDARRVHIDVRADDTGS
 MSVVLPRVLLTGPTRIAVSDQSHQGIWSGAPWTKLVLLPARADKTPVEG
 >MycoTub_F1_00015
 ELDVMRRRALATAAESAEQVLVTAAVLEDIPAGWDARRVHIDVRADDTGS
 MSVVLPRVLLTGPTRIAVSDQSHQGIWSGAPWTKLVLLPARADKTPVEG
 GSLPLRRVGWPDGDEAGRVRGPTFVIHNRWVRPLVSWQRNAAGRCGSPRV
 WCSASGLRWSATRSPNMHARLRTTGCALRSRRRGRRSGSCRPSFWPRSPQ
 >MycoTub_F1_00016
 GSLPLRRVGWPDGDEAGRVRGPTFVIHNRWVRPLVSWQRNAAGRCGSPRV
 WCSASGLRWSATRSPNMHARLRTTGCALRSRRRGRRSGSCRPSFWPRSPQ
 RLATMWCDRRSPGRRHHRGARGLAAILPVGV RATPTDNTSIGPEQGAPVPA
 EPRGRSGSVCR TSGRTRKKDVRRTARLDGKRGQKIGPGARSRYSGVRPPA
 >MycoTub_F1_00017
 RLATMWCDRRSPGRRHHRGARGLAAILPVGV RATPTDNTSIGPEQGAPVPA
 EPRGRSGSVCR TSGRTRKKDVRRTARLDGKRGQKIGPGARSRYSGVRPPA
 ATQPHARNPGEYS DRGCPEKEGPRRI RRCVYHHSRRAGGRPQTSRHVHWL
 DRARFTPSHLGGGRQGRRGDGR LCNHSERSAAGWRCRGRRRRPRHSGRH
 >MycoTub_F1_00018
 ATQPHARNPGEYS DRGCPEKEGPRRI RRCVYHHSRRAGGRPQTSRHVHWL
 DRARFTPSHLGGGRQGRRGDGR LCNHSERSAAGWRCRGRRRRPRHSGRH
 PRLRHTDRRRGDDTTTCRRQVRLGRVCDIWW SARRRRVGGRAIHPARSRD
 QARRVRVVSGL EVGTPGPQARGADQEDGVNGAVLGRPRCFRNHGIRLNR
 >MycoTub_F1_00019
 PRLRHTDRRRGDDTTTCRRQVRLGRVCDIWW SARRRRVGGRAIHPARSRD
 QARRVRVVSGL EVGTPGPQARGADQEDGVNGAVLGRPRCFRNHGIRLNR
 RPPAARDGVPQQGADHQPDREGDPRRGRRRSGQRRRRGAEVGKTRSR IH
 CTAQSEPHLSLSGWPGGLRETHQPHQERDSQHRGLFRQGHRARGGDRDAM
 >MycoTub_F1_00020
 RPPAARDGVPQQGADHQPDREGDPRRGRRRSGQRRRRGAEVGKTRSR IH
 CTAQSEPHLSLSGWPGGLRETHQPHQERDSQHRGLFRQGHRARGGDRDAM
 ERRVFGVGAHLRQHHQHPRGRHPRRGLPQRADVGGEQVRQGPQATEGQGP
 QPHRRYPGRPGRCDLGEGQRTAVRGPDQDQVGQHRGQIVCAEGLRTADPL
 >MycoTub_F1_00021
 ERRVFGVGAHLRQHHQHPRGRHPRRGLPQRADVGGEQVRQGPQATEGQGP
 QPHRRYPGRPGRCDLGEGQRTAVRGPDQDQVGQHRGQIVCAEGLRTADPL
 VSQPHRRRESRCEQGCVLGASPYRGTGTRVGAAERHHRHWIARQAGRLPFH
 GSAQVRTVCRRRLGRRFCCKRSRFDVPGD TSAARQDHQCGESAHRPGAKE
 >MycoTub_F1_00022
 VSQPHRRRESRCEQGCVLGASPYRGTGTRVGAAERHHRHWIARQAGRLPFH
 GSAQVRTVCRRRLGRRFCCKRSRFDVPGD TSAARQDHQCGESAHRPGAKE
 HRSSGDHHGAGHRDPRRVRYRQAALPQDRADGRRRCRPTYFHAVVDVVVP
 VHAAHRERACVFGTTAAVQTQVAAQPGIRILRPRARRSAGGGAEGREED