UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

SECRETARIA
DE
PÓS-GRADUAÇÃO
I. B.

ITARAJU JUNIOR BARACUHY BRUM

"FERRAMENTAS DE BIOINFORMÁTICA PARA PROTEÔMICA"

Este exemplar corresponde à redação final da tese defendida pelo(a) candidato (a) Itaraju Junior Baracuhy Brum

e aprovada pela Comissão Julgadora.

Dissertação apresentada ao Instituto de Biologia para obtenção do Título de Mestre em Biologia Funcional e Molecular, na área de Bioquímica.

Orientador: Prof. Dr. Eduardo Galembeck

Campinas, 2007

**FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP**

B834f

Brum, Itaraju Junior Baracuhy
Ferramentas de bioinformática para proteômica / Itaraju Junior Baracuhy Brum. – Campinas, SP: [s.n.], 2007.

Orientador: Eduardo Galembeck.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Bioinformática. 2. Proteômica. 3. Eletroforese. 4. Redes e vias metabólicas. I. Galembeck, Eduardo. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

(rcdt/ib)

Título em inglês: Bioinformatics tools for proteomics.
Palavras-chave em inglês: Bioinformatics; Proteomics; Electroforesis; Metabolic networks and pathways.
Área de concentração: Bioquímica.
Titulação: Mestre em Biologia Funcional e Molecular.
Banca examinadora: Eduardo Galembeck, José Camillo Novello, Cristina Pontes Vicente.
Data da defesa: 23/07/2007.
Programa de Pós-Graduação: Biologia Funcional e Molecular.

Campinas, 23 de Julho de 2007.

BANCA EXAMINADORA

Prof. Dr. Eduardo Galembeck (Orientador)

Assinatura

Prof. Dr. José Camillo Novello

Assinatura

Profa. Dra. Cristina Pontes Vicente

Assinatura

Prof. Dr. Sérgio Marangoni

Assinatura

Profa. Dra. Mônica Andréa Pickholz

Assinatura

Aos meus pais, Itaraju e Uiara.

Ao meu orientador, Eduardo, pelo imenso apoio, confiança e companheirismo.

À Unicamp, pelas oportunidades que me tem dado e o sonho que representou.

Ao meu mestre na vida, Daisaku Ikeda, pelo exemplo humano em abrir novos caminhos com imensa coragem.

Agradecimentos

Aos meus pais, Itaraju e Uiara, pelo apoio incondicional, pelo exemplo de valorização dos estudos, pela confiança e expectativa depositados e pela participação em todos esforços.

A meus irmãos, Itanara e Italo, que sempre estão próximos e compartilharam comigo as alegrias e esforços na vida. Agradeço meus grandes amigos da BSGI pela confiança e inúmeros incentivos.

A meu orientador, Dr. Eduardo Galembeck, pelas inúmeras oportunidades que me proporcionou, pelo acompanhamento e apoio ao longo de toda formação universitária, pela consideração e companheirismo.

A meu amigo Renato Alas Martins, que me apresentou ao Prof. Eduardo Galembeck e que vem contribuindo com incentivos, opiniões e mostrando novos caminhos.

Aos amigos do Laboratório de Tecnologia Educacional, IB-Unicamp, Daniela Kiyoko Yokaichiya, Gabriel Gerber Hornink, Renato Milani, Anderson Martins, Daniel Perez, Isabel Settin, Francisco Neto, Carlos Eduardo Santoro, Mário Sarraipa, Eduardo Kimura, Elaine Oliveira, Gesivaldo Santos, Bianca Rossi, pela amizade, apoio e agradável convivência.

Aos companheiros do Laboratório de Proteomica, IB-Unicamp, Marcus Bustamante Smolka, Daniel Martins, Flávia Vischi Winch, Bruno de Oliveira e Prof. José Camillo Novello por todo apoio na área de proteômica.

Aos professores participantes da banca de qualificação, Dr. José Camillo Novello, Dr. Gonçalo Guimarães Pereira e Dr. Marcelo Brocchi pela disponibilidade em avaliar este trabalho.

Aos professores participantes da banca examinadora, Profa. Dra. Cristina Pontes Vicente e Prof. Dr. José Camillo Novello pela análise da dissertação, comentários e tempo empregado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa concedida para realização deste projeto de mestrado.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelas bolsas de iniciação científica concedidas e que possibilitaram os trabalhos iniciais que culminaram neste mestrado.

Resumo

A área de proteômica visa estudar um conjunto completo de proteínas expressas por um organismo ou tecido numa dada situação, através da identificação e quantificação. Apesar de limitações nas técnicas disponíveis, vem se aumentando o volume de informações oriundos desta área, situação que exige o emprego de ferramentas computacionais para permitir o uso eficiente de dados disponíveis, além de buscar-se novas formas de análise destes.

Este projeto visa o desenvolvimento de ferramentas de bioinformática para aplicação em proteômica. Estas ferramentas abrangem as seguintes aplicações: Cálculo Teórico de Ponto Isoelétrico e Peso Molecular de seqüências de aminoácidos, eletroforesebidimensional teórica, digestão teórica e simulação de eletroforese e identificação de peptídeos, ferramenta para análise de Vias Metabólicas a partir de dados de proteômica.

Abstract

The proteomics field aims to study sets of proteins expressed in a cell or tissue, according to a specific situation, through protein identification and quantification. Though technical limitations do exist, the amount of information derived from this field increases each day. And so, there is a need for employing computational tools that enable efficient analysis of data.

This project aims developing bioinformatics tools for application in proteomics. The tools here presented comprehend the following tasks: theoretical computation of isoeletric point and molecular weight of aminoacid sequences, theoretical two-dimensional electrophoresis, theoretical triptic digestion and electrophoresis simulation for peptide identification, and analysis of metabolic pathways with proteomics data.

Sumário

Sumário vii

Lista de Figuras viii

Lista de Tabelas ix

1 Introdução: proteômica e bioinformática 1
- 1.1 A área de Proteômica . . . 1
  - 1.1.1 Eletroforese de duas dimensões (2D) . . . 1
  - 1.1.2 Identificação de proteínas . . . 2
  - 1.1.3 Espectrometria de massa . . . 3
  - 1.1.4 Espectrometria de massa com ICAT . . . 5
- 1.2 Bioinformática em proteômica . . . 9
  - 1.2.1 Mapa 2D teórico . . . 9
  - 1.2.2 O servidor ExPASy . . . 9
- 1.3 Bancos de dados sobre vias metabólicas . . . 10
  - 1.3.1 Banco de Dados KEGG . . . 10
  - 1.3.2 Banco de dados BioCyc . . . 11
- 1.4 Objetivos . . . 14
  - 1.4.1 Ferramentas para géis 2D . . . 14
  - 1.4.2 Identificação de peptídeos . . . 15
  - 1.4.3 Ferramenta de análise por vias metabólicas . . . 15

2 Ferramentas para géis 2D 16
- 2.1 Cálculo de pI e MW e Mapa 2D teórico . . . 16
  - 2.1.1 Métodos . . . 16
  - 2.1.2 Resultados . . . 20
  - 2.1.3 Testes . . . 22
- 2.2 Estatísticas de N-terminais . . . 27

3 Digestão teórica e identificação de peptídeos 32
- 3.1 Métodos . . . 32
  - 3.1.1 Visualização de peptídeos . . . 32
  - 3.1.2 Avaliação teórica da estratégia . . . 32
  - 3.1.3 Desenvolvimento de ferramenta de busca . . . 33

3.2 Resultados 33
3.2.1 Obtenção de peptídeos 33
3.2.2 Formas de registro de peptídeos 35
3.2.3 Resultados de visualização 38
3.2.4 Resultados de consultas estatísticas 39
3.2.5 Ferramenta de Busca 43

4 Vias metabólicas **54**
4.1 Obtenção de uma representação para os dados 54
4.2 Ferramenta de consulta: ECPath 56
4.2.1 Interface para a ECPath 58

5 Conclusão **61**
5.1 Discussão geral das ferramentas 61
5.2 Sugestões para desenvolvimento futuro 62
5.2.1 Banco de dados de proteoma 63
5.2.2 Mapa 2D teórico e estatísticas N-terminal 65
5.2.3 Digestão teórica e identificação de peptídeos 66
5.2.4 Vias metabólicas 66

A Produção bibliográfica **68**
A.1 Registros de programas de computador 68
A.2 Artigo completo publicado em periódico 68
A.3 Resumos publicados em anais de eventos 68

Referências Bibliográficas **71**

Lista de Figuras

1.1 Representação esquemática de um gel 2D. 2
1.2 Esquema: *Peptide Mass Finger-printing* 4
1.3 Estrutura do reagente ICAT 5
1.4 Identificação de peptídeos com ICAT 7
1.5 Via metabólica da KEGG 12
1.6 Vias metabólicas no BioCyc 13

2.1 Resultados de cálculo de pI e MW 21
2.2 Mapa 2D teórico gerado 22
2.3 Testes pI e MW com *Escherichia coli* 25
2.4 Testes pI e MW com *Xylella fastidiosa* 26
2.5 Comparação em Mapas 2D teóricos 27
2.6 Comparação em Mapas 2D teórico e experimental 28
2.7 Tabela com listagem N-terminal 29
2.8 ORFs com N-terminal “MIF” 31

3.1 Gel 2D teórico de peptídeos 38
3.2 Gel 2D teórico de peptídeos em escala menor 39
3.3 Quantidade de proteínas pelo número de cisteínas 40
3.4 Proteínas em faixas de MW 41
3.5 Contagem de regiões: MW < 20kDa 44
3.6 Contagem de regiões: 20kDa < MW < 50kDa 45
3.7 Contagem de regiões: MW > 50kDa 45
3.8 Contagem de regiões: Geral 46
3.9 Formulário da FindPep 49
3.10 Página de resultado de busca de peptídeos 50

4.1 Estrutura dos dados sobre vias metabólicas 55
4.2 Formulário de submissão da ECPath 58

5.1 Página inicial do Banco de dados 64

Lista de Tabelas

2.1 Valores de pK para aminoácidos ionizáveis 18
2.2 Valores de MW para aminoácidos 20
2.3 Valores de pI calculados e valores obtidos de *sites* externos 23
2.4 Valores de pK para os 20 aminoácidos 24

3.1 MW monoisotópicas dos aminoácidos 34
3.2 MW do reagente ICAT 34
3.3 Digestão teórica de uma ORF 35
3.4 Estrutura da tabela "proteínas" 37
3.5 Estrutura da tabela "peptídeos" 37
3.6 Número de peptídeos e cisteínas 40
3.7 Tamanho de faixas de MW para peptídeos 42
3.8 Exemplos de peptídeos muito comuns 44
3.9 Proteínas utilizadas em testes. 52
3.10 Peptideos originados de digestão teórica de uma proteína 53

4.1 Tabelas de dados para vias metabólicas 57
4.2 Exemplo de resultado da ECPath - lista de reações 59
4.3 Exemplo de resultado da ECPath - vias metabólicas 60

Capítulo 1

Introdução: proteômica e bioinformática

1.1 A área de Proteômica

A Proteômica (do inglês, *Proteomics*) é um novo campo de estudo e tecnologia que se propõe a analisar de forma global o conjunto de proteínas expresso numa célula ou tecido, isto é, o proteoma (Wilkins e Hochstrasser, 1997). Na era pós-genômica, o surgimento da Proteômica está diretamente relacionado à necessidade de se investigar o controle da expressão gênica em escala global. Importantes informações podem ser geradas, como:

- Quais proteínas são expressas.
- Níveis de expressão destas proteínas.
- Modificações pós-transcricionais.
- Respostas das células a diferentes situações/tratamentos.
- Diferenças entre linhagens.

Devido à natureza dinâmica do Proteoma (ele se altera frente a diferentes condições e estímulos), seu estudo representa uma forma de procurar possíveis funções das proteínas e uma forma global de investigar processos em sistemas vivos para melhor entender o funcionamento de uma célula ou tecido ao nível molecular.

1.1.1 Eletroforese de duas dimensões em gel de poliacrilamida (2D)

A eletroforese de duas dimensões em gel de poliacrilamida é o método disponível mais poderoso para separar misturas complexas de proteínas (Wilkins e Hochstrasser, 1997; Westermeier, 1997). Logo, a 2D representa a base da tecnologia da Proteômica.

A primeira dimensão da 2D é a focalização isoelétrica ("IEF"), na qual as proteínas são separadas em um gradiente de pH até alcançarem a posição estacionária onde a carga total é zero. O pH no qual a proteína tem carga total zero é chamado de ponto isoelétrico (pI). Na segunda dimensão, as proteínas separadas pela IEF são novamente separadas por eletroforese do tipo SDS-PAGE. Essa separação é baseada no peso molecular (MW) das proteínas.

O resultado da separação é um perfil onde proteínas são representadas por pontos, ou *spot* (Figura 1.1), sendo que quanto maior o *spot* maior a quantidade presente da proteína na amostra. Segundo um sistema cartesiano, da esquerda para a direita, há um aumento do pI e de baixo para cima um aumento do peso molecular. A alta resolução da 2D resulta do fato da primeira e segunda dimensões serem baseadas em parâmetros independentes (pI e peso molecular das proteínas).

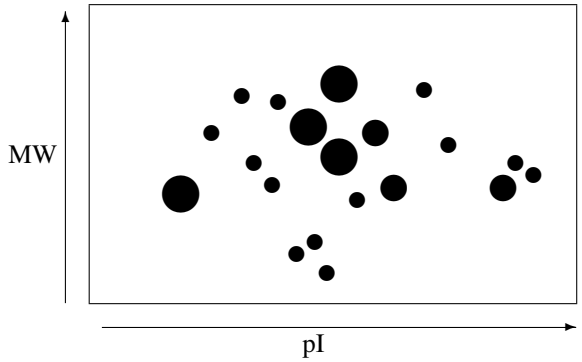


Figura 1.1: Representação esquemática de um gel 2D.

As imagens de géis 2-D podem ser digitalizadas para subseqüentes análises quantitativas e qualitativas utilizando-se programas computacionais especializados. Dessa forma, para cada *spot*, é possível calcular o pI e o peso molecular aparente (MW), quantificá-lo e compará-lo automaticamente em diferentes géis.

1.1.2 Identificação de proteínas

Para organismos com genoma inteiramente seqüenciado, a identificação de proteínas se baseia em comparar os dados experimentalmente obtidos pela análise dos mapas 2D com as predições de *Open Reading Frames* ("ORFs") geradas pela análise das informações do genoma. A identificação resulta no conhecimento da seqüencia completa da proteína identificada.

Após a separação por 2D, é geralmente necessário obter dados adicionais sobre cada proteína individualmente, para então relacioná-la à específica ORF. Dentre as metodologias de análise pós-separação mais utilizadas para a identificação estão:

- Análise da composição de aminoácidos.
- Seqüenciamento N-terminal.

- Espectrometria de massa.

Estas análises em associação com os dados de pI e a MW estimados por análise da imagem do gel, possibilita a identificação inequívoca da ORF específica.

De forma geral, as características das proteínas são confrontadas contra um banco de dados de ORFs, onde os dados teóricos destas características experimentalmente analisadas são calculadas a partir da seqüencia da ORF. Por exemplo, é possível calcular o pI e o MW, e prever o N-terminal a partir da seqüencia nucleotídica de uma ORF e confrontar estes dados com o pI, a MW e o N-terminal experimentalmente determinados (Urquhart *et al.*, 1998; Wilkins *et al.*, 1998).

A seqüencia de proteínas, mesmo que de apenas alguns aminoácidos, é consideravelmente específica. Por exemplo, existem 8.000 combinações possíveis de seqüências de 3 aminoácidos, 160.000 combinações de seqüências de 4 aminoácidos, e 3.200.000 combinações de seqüências de 5 aminoácidos (Wilkins e Gooley, 1997).

Recentemente, pequenos pedaços de seqüências da extremidade N-terminal de proteínas têm sido propostos como atributos para identificação de proteínas. Sua especificidade é surpreendente, especialmente em organismos com genomas pequenos. Por exemplo, em *E. coli*, por volta de 60% das proteínas têm um único N-terminal de 4 aminoácidos de extensão.

Mesmo quando o N-terminal não é único para uma proteína, relativamente poucas proteínas compartilham o mesmo. Nestes casos, os dados de pI e MW estimados devem ser adicionalmente utilizados para identificação.

Isto mostra que o N-terminal de aproximadamente 5 aminoácidos, quando determinado analiticamente a partir de proteínas de géis 2D, deve ser atributo poderoso para identificação. A aplicação ideal de seqüenciamento N-terminal é para identificação de proteínas de organismos de proteoma pequeno, que tenham genoma já conhecido (Wilkins e Gooley, 1997).

1.1.3 Espectrometria de massa

Hoje, a espectrometria de Massa (MS) é a técnica mais utilizada para identificação de proteínas de géis 2DE (Gygi e Aebersold, 2000). Basicamente, para identificação de proteínas por MS, os *spots* são extraídos do gel 2D, submetidos a uma digestão proteolítica *in-gel* e a massa dos peptídeos resultantes analisados no espectrômetro, de onde se obtém suas massas. Os valores de massa são então confrontados contra uma lista, gerada por computador, obtida pela digestão *in-silico* de um banco de dados de proteínas ou banco de dados traduzido de nucleotídeos usando-se a mesma "enzima" utilizada na digestão experimental (Lahm e Langen, 2000). Esta técnica é conhecida como *Peptide Mass Fingerprinting*, Figura 1.2, e é uma ferramenta aplicável na análise de proteínas de organismos que já tenham seu genoma seqüenciado. Por análise de imagem do perfil 2DE, é possível quantificar os níveis de expressão de proteínas baseada na intensidade de *spots* (Pleissner *et al.*, 2001).

Entretanto, a 2DE-MS, apesar do reconhecimento geral de sua utilidade para a análise global de proteoma, apresenta deficiências, herdadas do uso combinado com a 2DE, que sugerem que ela não seja realmente tão global (Gygi e Aebersold, 2000). Classes de proteínas como as muito ácidas ou básicas, muito pequenas ou grandes,

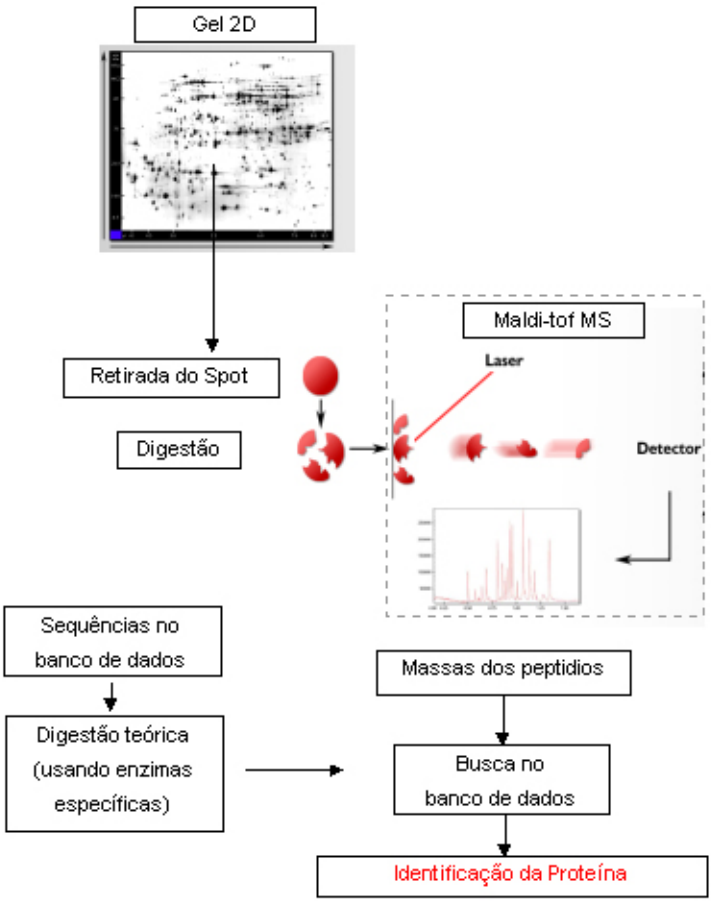


Figura 1.2: Esquema de funcionamento da técnica *Peptide Mass Finger-printing*.

proteínas muito hidrofóbicas, proteínas pouco abundantes não são analisadas pela 2DE-MS.

Além disso, a quantificação não é muito precisa devido a problemas de linearidade das técnicas de visualização (Patton, 2001) e, principalmente pela baixa reprodutividade da separação na 2DE, o que dificulta a análise de imagem comparativa entre géis (Smolka *et al.*, 2001).

1.1.4 Espectrometria de massa com ICAT

Uma técnica para quantificação precisa e simultânea identificação de seqüencias de proteínas em misturas complexas, baseada na etiquetagem de proteínas com reagentes do tipo *Isotope Coded Affinity Tag* (ICAT), foi descrita por Gygi *et al.* (1999). O método é baseado em dois princípios. Primeiro, uma seqüência curta e contígua de aminoácidos de uma proteína (5–25 resíduos) contém informação suficiente para identificar uma proteína. Segundo, pares de peptídeos marcados com reagentes ICAT leve e pesado (ver discussão abaixo), são quimicamente idênticos e, por isso, servem como parâmetros para quantificação precisa num experimento de MS (Gygi *et al.*, 1999). O reagente ICAT consiste de três componentes funcionais, conforme a Figura 1.3. O primeiro é um grupo reativo seletivo para as sulfidrinas das cadeias laterais dos resíduos de cisteína reduzida. O segundo componente é um ligante de etileno glicol que pode estar ligado a 8 deutérios (forma isotopicamente pesada) ou a 8 hidrogênios (forma isotopicamente leve) e que é a base da quantificação por espectrometria de massa. O terceiro componente é uma biotina que possibilita que os peptídeos etiquetados com ICAT sejam isolados por cromatografia de afinidade com avidina (Smolka *et al.*, 2001).

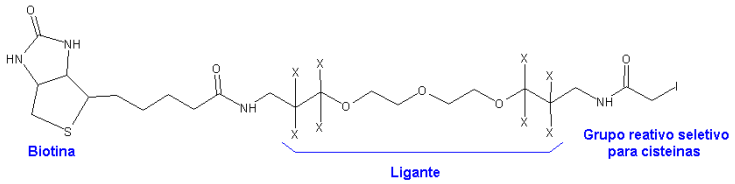


Figura 1.3: Estrutura do reagente ICAT. “X” indica as posições de deutérios (forma pesada) ou hidrogênios (forma leve).

Na análise com este reagente, primeiramente, as proteínas são marcadas com o ICAT, que se liga a resíduos de cisteína.. Para duas amostras sendo comparadas, uma é marcada com forma leve (não deutérica) e a outra com a forma pesada (deutérica) do reagente. Ambas amostras são combinadas e, juntas, digeridas enzimaticamente com tripsina, resultando numa mistura complexa de peptídeos. Após passos seqüenciais de cromatografia líquida, a identificação de proteínas é obtida por análise de MS/MS dos peptídeos resultantes. A quantificação é possível porque peptídeos marcados com ICAT leve e pesado são exibidos como picos distintos nas corridas de MS (mantendo uma diferença de 8Da), e as intensidades relativas destes picos são proporcionais à abundância relativa da proteína original nas amostras iniciais (Smolka *et al.*, 2001). Esta abordagem é mais global do que a 2DE-MS, pois permite detecção e quantifica-

ção de classes de proteínas não observadas na 2DE como as muito básicas, hidrofóbicas, pequenas; podendo ser usada para virtualmente todo o proteoma. A técnica do ICAT permite o uso de grande quantidade de amostra, permitindo detecção de proteínas pouco abundantes, e permite uma quantificação muito precisa (Gygi *et al.*, 1999).

Uma abordagem para identificação de proteínas, descrita abaixo, prevê o uso de uma série de critérios de seleção que, obtidos experimentalmente, podem ser usados para a suficiente restrição de buscas em bancos de dados genômicos. A estratégia pode ser descrita em seis passos (Figura 1.4):

1. **Etiquetagem de proteínas**. Uma amostra de proteínas é etiquetada com reagentes ICAT. Um reagente ICAT liga-se, por ligação covalente, a cada resíduo de cisteína presente em todas as proteínas da amostra.

2. **Separação por SDS-PAGE** (Opcional). As proteínas presentes na amostra, já etiquetadas, são separadas segundo seu peso molecular (MW) através de eletroforese de SDS-PAGE. E seleciona-se, da amostra resultante, as proteínas dentro de uma faixa de MW a ser estudada. Espera-se que, para qualquer faixa estudada, seja selecionada potencialmente um grande número de proteínas.

3. **Digestão por tripsina**. A amostra resultante é submetida a uma digestão, de onde se obtém uma série de peptídeos resultantes da clivagem das proteínas em resíduos específicos.

4. **Seleção por afinidade por avidina**. O conjunto de peptídeos selecionados até o momento e que contém o resíduo de cisteína mantém a etiquetagem com o ICAT que, por sua vez, possui o grupo biotina. Assim, os peptídeos etiquetados podem ser isolados por cromatografia de afinidade de avidina. Ou seja, pode-se obter uma nova amostra, agora com somente peptídeos que contém resíduo de cisteína.

5. **Separação por pI**. O conjunto de peptídeos é submetido a uma separação por focalização isoelétrica, ou seja, por Ponto Isoelétrico. De onde podem ser separados os peptídeos que podem ser encontrados numa determinada faixa de pI.

6. **MS**. Os peptídeos selecionados (dentro de uma faixa de pI) são submetidos a espectrometria de massa do tipo Maldi-ToF, onde seus pesos moleculares podem ser medidos com alta precisão.

Em resumo, a seqüência de passos apresentada fornece o seguinte conjunto de informações sobre um peptídeo em estudo ao final do processo:

- Têm-se uma faixa de MW para a proteína que o originou (passo 2).
- Sabe-se, pelo reagente da digestão, os resíduos que marcam o início e fim de peptídeos (passo 3).
- os peptídeos contém pelo menos uma cisteína, devido à etiquetagem com ICAT (passos 1 e 4).

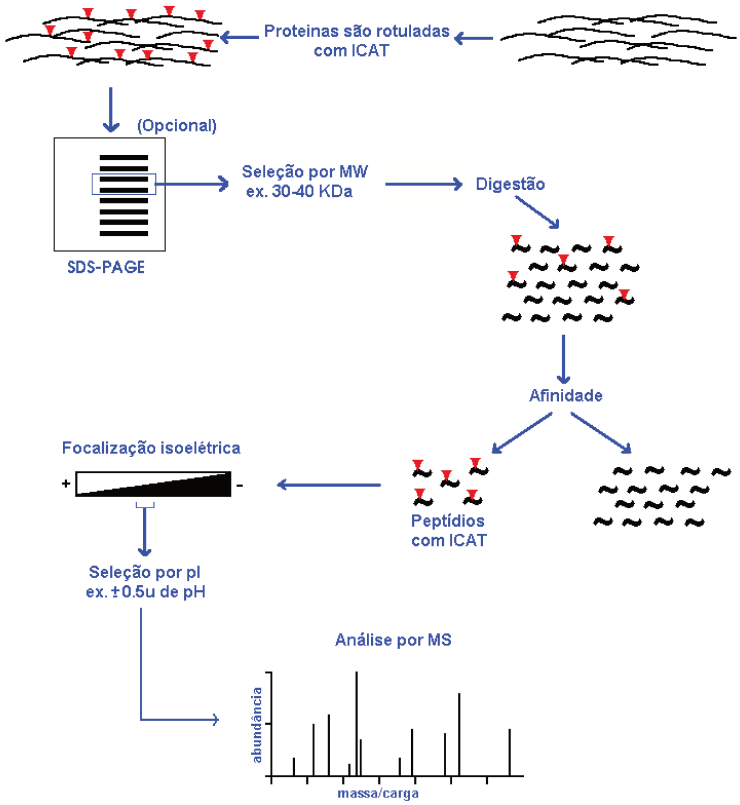


Figura 1.4: Esquema geral da proposta de estratégia para identificação de proteínas com aplicação de ICAT. Estão indicados os passos experimentais para obtenção de parâmetros de busca contra um banco de dados de proteínas.

- têm-se uma faixa de pI para o peptídeo (passo 5).
- conhece-se a massa molecular do peptídeo com precisão de 70 ppm (passo 6).

Essas informações, conjugadas, podem ser confrontadas *in-silico* contra uma base de dados de seqüências de aminoácidos, servindo como critérios de seleção para correlacionar um peptídeo estudado com uma ou mais proteínas presente no banco de dados, assim, identificando-a.

Quanto menor for o número de proteínas com as quais um peptídeo puder ser correlacionado, mais seletivo é o processo. No caso da seletividade for suficiente para obter-se uma única proteína correlacionada, do banco de dados, pode-se concluir que a abordagem é suficiente para a identificação de proteínas. A estratégia, então, baseia-se em obter dados experimentais que possam ser comparados com dados teóricos para restringir um banco de dados de proteínas de forma a resultar em identificação única e inequívoca.

Cabe ressaltar alguns aspectos positivos da abordagem. Primeiramente a SDS-PAGE é mais global que a 2DE (permitindo a visualização de maior quantidade de proteínas hidrofóbicas e grandes, por exemplo) e é possível "correr" rapidamente várias amostras em paralelo, variando a quantidade empregada a fim de obter melhor visualização de proteínas pouco abundantes. Em segundo lugar, a separação de peptídeos que contenham cisteínas é muito seletiva, por exemplo, a digestão teórica por tripsina do proteoma completo da levedura *Saccharomyces cerevisiae* (6.113 proteínas) produz 344.855 peptídeos, no entanto, somente 30.619 destes apresentam resíduo de cisteína (Gygi *et al.*, 1999). Vê-se também que a medida de pI realizada no passo 5 pode ser muito melhor correlacionada com previsões teóricas uma vez que, sendo a separação por pI realizada sobre peptídeos, espera-se que as cargas dos peptídeos estejam melhor expostas (algo que é assumido no cálculo teórico, ver 2) pela impossibilidade de formação de estruturas terciárias com pequenos peptídeos. Finalmente, a medida de MW de peptídeos por MS é muito precisa e facilmente correlacionada com cálculos teóricos.

A maior vantagem desta técnica reside no fato de, ao contrário da estratégia inicialmente proposta por Gygi *et al.* (1999), haver a possibilidade de identificação de proteínas sem o uso de MS/MS (por espectrometria de CID), permitindo maior rapidez e simplicidade de análise e interpretação dos resultados de identificação e quantificação. Para o uso da estratégia é importante avaliar qual tamanho que as faixas de MW e pI devem ter nos passos 2 e 5. Faixas maiores levam a uma maior quantidade de proteínas na busca do banco de dados, resultando numa menor seletividade. Há, entretanto, a possibilidade de o aumento da faixa de algum desses parâmetros não diminuir significativamente a seletividade do método, de modo a permitir faixas mais abrangentes de busca e até a eliminação de algum dos passos descritos. Isso pode ser particularmente útil para o caso da faixa de MW. A SDS-PAGE é trabalhosa e se fosse possível a eliminação do passo 2, sem detrimento da estratégia, conseguiría-se uma melhoria significativa no uso prático desta estratégia.

1.2 Bioinformática em proteômica

Nota-se, que a utilização de ferramentas computacionais é essencial em um projeto de análise de Proteomas, principalmente na etapa de identificação de proteínas. Programas para cálculo de pI e MW, construção de mapa 2D teórico e análise de resíduos N-terminais são alguns exemplos de ferramentas necessárias.

O desenvolvimento de ferramentas computacionais envolve a confecção dos programas que processarão os dados e da interface do usuário, sendo necessário o conhecimento de algumas linguagens de programação. Também é necessário conhecimento de bancos de dados, para a estruturação das informações que serão processadas. Para o desenvolvimento dos algoritmos é necessário, além de conhecer as ferramentas computacionais, conhecer o assunto sobre o qual se está trabalhando (Bioquímica, no caso) e converter os problemas bioquímicos em problemas matemáticos que possam ser processados pelos computadores e retornar um dado que possa ser compreendido pelos bioquímicos. Esta problemática muitas vezes requer um trabalho interdisciplinar, de biólogos e programadores, para que se obtenham as melhores soluções.

1.2.1 Mapa 2D teórico

Um mapa 2D teórico é construído após o cálculo de pI e MW teórico de todas as ORFs. As ORFs são então "plotadas" em coordenadas cartesianas similares ao do mapa de referência 2D real. Desta forma, se obtém a localização teórica da ORF em um mapa 2D, experimental.

O mapa teórico é muito útil no processo de identificação de proteínas. É possível saber a densidade de ORFs por área, e assim prever o nível de dificuldade para identificar um *spot* presente numa determinada região, bem como levantar uma lista de ORFs candidatas.

1.2.2 O servidor ExPASy

Como uma das principais referências para ferramentas computacionais na área de proteômica, tem-se o servidor ExPASy[1] (*Expert Protein Analysis System*). É um servidor dedicado à análise de seqüências e estruturas de proteínas e géis 2D, mantendo um grande volume de informações em bancos de dados e também ferramentas de bioinformática para suporte à análise.

As ferramentas podem ser acessadas via Internet e estão distribuídas em diferentes áreas e também pode-se obter referências para ferramentas em outros servidores (Gasteiger *et al.*, 2003, 2005). Estão representadas áreas como:

- Identificação e caracterização de proteínas
- Tradução de seqüências de nucleotídeos em aminoácidos e análise estatística
- Buscas de padrões e perfil
- Predição de modificações pós-transcricionais

[1] http://www.expasy.org, Jul/2005.

- Predição de topologia
- Análise de estrutura primária
- Predição de estrutura secundária
- Estrutura terciária
- Alinhamento de seqüências
- Análise de textos biológicos

Entre os bancos de dados disponíveis, encontramos o Swiss-Prot, que mantém dados processados manualmente sobre seqüências de proteínas. O Swiss-Prot visa manter, mesmo que a custo de um volume menor de dados, informações de alto nível de anotação com dados como descrição de função da proteína, domínios estruturais, modificações pós-transcricionais, variações, etc. As entradas são analisadas para se ter um nível baixo de redundância de seqüências e também manter-se referências a outras fontes de informações e bancos de dados sobre as entradas. Já com uma outra abordagem, o TrEMBEL é um banco de dados no ExPASy que procura manter o maior volume de seqüências disponíveis sobretudo anotadas automaticamente por computador. Estas são entradas que não integram o Swiss-Prot (Boeckmann *et al.*, 2003).

O banco de dados de géis 2D é o SWISS-2DPAGE. Ele mantém informações sobre proteínas identificadas em mapas de referência 2D e SDS-PAGE. Atualmente, encontram-se 36 mapas de vários tecidos do homem e rato e organismos como *Arabidopsis thaliana*, *Dictyostelium discoideum*, *Escherichia coli*, *Saccharomyces cerevisiae* e *Staphylococcus aureus (N315)*. Este banco de dados permite a visualização dos mapas armazenados e, a partir dela, o acesso às proteínas identificadas no mapa, referenciando dados anotados (Hoogland *et al.*, 2004).

1.3 Bancos de dados sobre vias metabólicas

Informações sobre vias metabólicas já estudadas podem ser obtidas na literatura. No entanto, além de dispersas e dispostas em diferentes formatos, com qualidade diferente de informação e descrição, estas prestam mais à consulta do que o uso para processamento computacional. O que também se observa em algumas versões eletrônicas delas. Assim, a existência de bancos de dados de vias metabólicas como a KEGG e BioCyc (Lindroos e Andersson, 2002) com dados de diversas vias, reações químicas, enzimas envolvidas é fonte importante para o desenvolvimento de novas ferramentas em bioinformática.

1.3.1 Banco de Dados KEGG

O KEGG (*Kyoto Encyclopedia of Genes and Genomes*) (Kanehisa, 1997; Kanehisa e Goto, 2000) pode ser visto como uma ferramenta *web* para análise de interações moleculares e baseada em seqüências de genomas completos (Lindroos e Andersson, 2002). Mantém uma base de dados de vias que é divida em uma seção de vias metabólicas e

outra de vias de regulação, sendo a primeira melhor organizada e mais completa que a segunda (Ogata *et al.*, 1998). Atualmente, estão disponíveis 269 vias metabólicas de referência[2]. Os diagramas destas vias, conforme Figura 1.5, apresentam retângulos para enzimas e representam os demais compostos que participam das reações químicas por elas catalisadas. KEGG incorpora numa mesma figura diferentes variantes de uma via, originando representações com diferentes caminhos alternativos segundo o que se conhece atualmente sobre ela.

A KEGG também contém o banco de dados LIGAND com informações sobre enzimas, compostos químicos e reações químicas. Contém também um banco de dados sobre genes, com nomes de genes, seqüências e informação de função sobre genes de organismos seqüenciados. Todos esses bancos, incluindo aquele sobre vias metabólicas possuem campos chave que permitem a interligação entre de informações sobre eles, o que permite o desenvolvimento de várias ferramentas de bioinformática baseadas nelas.

Combinando-se os dados sobre organismos seqüenciados e as vias metabólicas, a KEGG reconstrói vias metabólicas específicas de organismos a partir das vias de referência. Na Figura 1.5 isto é mostrado pelos quadros de enzima marcados.

Atualmente, há 963.865 genes de 317 organismos seqüenciados, 6.475 reações químicas envolvendo 12.893 compostos nos bancos de dados[3]. Estes dados estão disponíveis via ftp para *download*[4].

1.3.2 Banco de dados BioCyc

Outro conjunto de bancos de dados com informações de genomas e vias metabólicas é BioCyc que mantém 601 vias metabólicas, 456 organismos, 2.458 enzimas e 5.273 reações enzimáticas, entre outros dados[5], obtidos da literatura (Karp *et al.*, 2002). O BioCyc apresenta algumas diferenças com a KEGG, como o fato de que as vias metabólicas serem previstas computacionalmente a partir dos genomas anotados, resultando em vias específicas para a espécie considerada e mais simples por registrar separadamente as vias variantes (Karp *et al.*, 2002). Ele também apresenta o recurso de se obter uma visão geral de todas as vias de um organismo cadastrado numa única operação. Bancos de dados e ferramentas do BioCyc podem também ser obtidos pela Internet.

No entanto, um aspecto importante no BioCyc é a possibilidade de colorir o diagrama de vista geral a partir de dados de experimentos de *Microarray* (Lindroos e Andersson, 2002), conforme a figura 1.6. Assim, têm-se uma integração das vias metabólicas de um organismo com os dados das diferenças de expressão gênica que são evidenciadas pela técnica de *Microarray*. Os dados podem ser fornecidos pelos usuários do MetaCyc.

Podemos observar também o surgimento de ferramentas computacionais que visam integrar resultados de Expressão de Gênica e Redes Metabólicas, fornecendo informações estatísticas como o programa Pathway Processor (Grosu *et al.*, 2002). Mesmo a KEGG desenvolveu o EXPRESSION, que mantém dados de expressão gênica e fornece ferramentas para estudo com vias metabólicas (Kanehisa *et al.*, 2002), no entanto,

[2] http://www.genome.ad.jp/kegg/kegg1.html, Jul/2005.
[3] http://www.genome.ad.jp/kegg/kegg1.html, Jul/2005.
[4] ftp.genome.ad.jp/pub/kegg, Jul/2005.
[5] http://biocyc.org/metacyc/release-notes.shtml, Jul/2005.

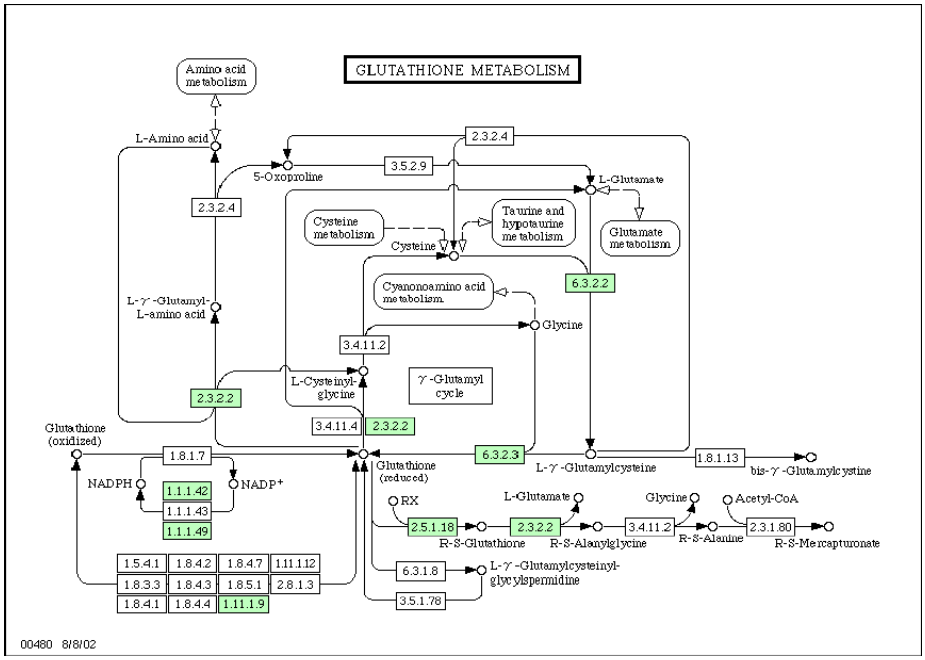


Figura 1.5: Exemplo de figura de via metabólica do banco de dados da KEGG (http://www.genome.ad.jp/kegg (Fev/2005)). Os quadros marcados evidenciam enzimas presentes no genoma da *Xylella fastidiosa*.

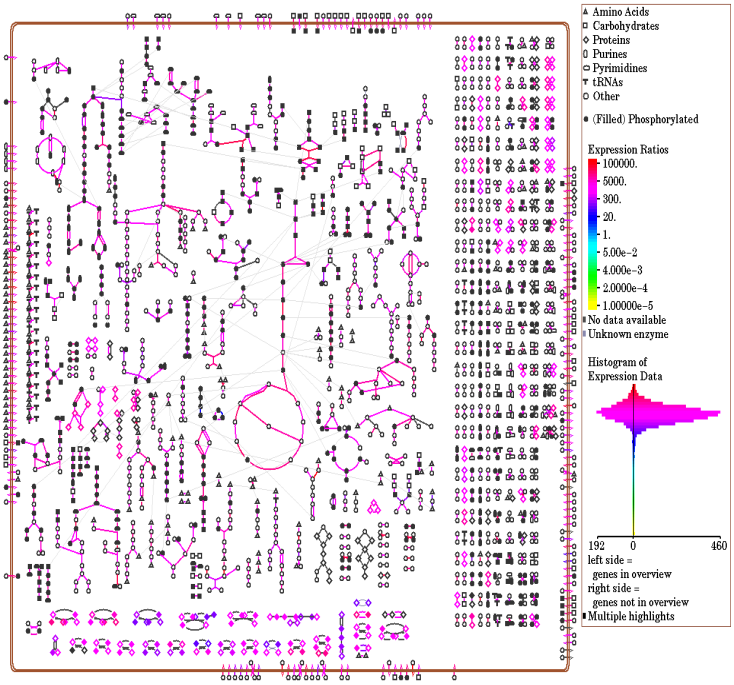


Figura 1.6: Exemplo de vista geral sob as vias metabólicas de *Escherichia coli* utilizando-se correlação com dados sobre expressão gênica (http://biocyc.org:1555-/expression.html (Jul/2005)).

ele não permite a submissão de dados além daqueles já coletados em seu banco de dados. Também se pôde notar o surgimento de esforços de integrar também informações da área de proteoma, indo além da expressão de genes (Ideker *et al.*, 2001).

1.4 Objetivos

Este trabalho visa o desenvolvimento de ferramentas de bioinformática para aplicação em proteômica. Estas ferramentas abrangem as seguintes aplicações: Cálculo Teórico de Ponto Isoelétrico e Peso Molecular de seqüências de aminoácidos, eletroforese-bidimensional teórica, digestão teórica e simulação de eletroforese e identificação de peptídeos, ferramenta para análise de Vias Metabólicas a partir de dados de proteômica. Os objetivos específicos destas ferramentas são descritos nas seções seguintes.

Devido à cooperação com demais projetos realizados no departamento, em várias momentos são utilizados como fonte de dados de referência o genoma ou proteoma do organismo *Xylella fastidiosa*.

1.4.1 Ferramentas para géis 2D

A primeira ferramenta visa calcular o pI e o MW teórico de todas as ORFs contidas num genoma (como exemplo, aproximadamente 2700 na *Xylella fastidiosa*) e, para isso, o programa deve poder realizar análise automática de seqüências de aminoácidos em larga escala. Apesar de existirem algumas ferramentas do gênero disponíveis para utilização pela Internet, elas não são adaptadas para utilização em escala genômica (só é possível submeter uma seqüencia de cada vez para as análises), o que inviabiliza seu uso no presente projeto. Quando o próprio programa é desenvolvido e utilizado internamente, além de torná-lo mais adaptado às reais necessidades, é ainda possível realizar as análises com maior velocidade.

Também será necessário utilizar um programa para criar automaticamente uma lista de N-terminais e um mapa 2D teórico contendo todos os genes de um genoma para comparação com o gel 2D obtido experimentalmente. Ferramentas desta natureza são de importância fundamental para identificação de proteínas e para construção de bancos de dados de proteoma.

1. Desenvolvimento de ferramentas de análise de seqüências:
    - Cálculo automático do pI (ponto isoelétrico) e MW (peso molecular) teóricos, com possibilidade de submissão de várias seqüências simultaneamente;
    - Obtenção de mapa 2D teórico;
    - Criação automática de lista de N-terminais teóricos com sistema específico de busca.
2. Adequação das ferramentas à interface para acesso *web*.

1.4.2 Identificação de peptídeos

- Avaliação teórica da estratégia de identificação de proteínas por peptídeos utilizando ICAT (Seção 1.1.4), verificando as faixas de parâmetros (pI e MW) que sejam suficientes para restringir uma busca para uma única proteína e também verificando a importância de usar SDS-PAGE.

- Desenvolvimento de ferramenta computacional para aplicação da estratégia.

Como descrito acima, objetiva-se avaliar teoricamente o método, avaliando a seletividade de cada passo, uma vez que quanto mais seletivos forem, maior é a restrição aplicada no banco de dados e maior a chance identificação única e inequívoca.

1.4.3 Ferramenta de análise por vias metabólicas

Partindo de dados da KEGG (Seção 1.3.1), que mantém um conjunto extenso de informações sobre vias metabólicas, é possível obter correlações com informações sobre expressão de genes obtidas do transcriptoma e proteoma. Podendo, deste modo, verificar a ocorrência de uma via metabólica com base nos genes efetivamente transcritos e proteínas produzidas efetivamente.

Dada a dinâmica do transcriptoma e proteoma, de modo que nem todas proteínas (ou mRNA) são expressas num mesmo instante numa célula e tecido ou variando no tempo de acordo com o desenvolvimento ou estado fisiológico da célula (Cahill *et al.*, 2001), a abordagem proposta pode possibilitar um estudo comparativo de vias metabólicas presentes em diferentes situações.

Para isso, deverá ser desenvolvido um Modelo, incluindo estrutura de dados (para armazenamento e eficiente recuperação de dados de vias metabólicas) e algoritmos para execução efetiva de comparações e visualização de resultados.

Capítulo 2

Ferramentas para géis 2D

2.1 Cálculo de pI e MW e Mapa 2D teórico

2.1.1 Métodos

Interface para entrada de dados

As ferramentas desenvolvidas fazem análise de seqüências de ORFs. Como têm-se o objetivo de se fornecer informações ao projeto de estudo da *Xylella fastidiosa*, a principal fonte de informações é o banco de dados de ORFs deste último projeto. Também são utilizados banco de dados *web* como o Expasy para testes de funcionamento.

Foi desenvolvida uma interface de entrada para tais bancos de dados através de um *site*, onde existem campos para que um usuário entre com dados de ORFs e possa submetê-los às análises desejadas.

Esta forma de implementação do projeto é especialmente útil pois disponibilizará para toda a comunidade acadêmica, interna à Unicamp, nacional e internacional, as ferramentas desenvolvidas. Os usuários devem preencher formulários simples, o que representa a entrada dos dados, e terão, pela própria Internet, os resultados esperados das ferramentas de análise.

Os dados de ORFs poderão ser encontrados de 2 modos:

1. Entrada manual (através de digitação ou métodos copiar/colar) das ORFs em uma caixa de texto presente no formulário.

2. Envio de um arquivo de dados de ORFs presente no computador do usuário. Este poderá ser executado através do uso de FTP (*File Transfer Protocol* - Protocolo de Transferência de Arquivos), sendo que a página exibirá instruções sobre este procedimento, ou *upload* do arquivo a partir de campos, da página, para designação do arquivo.

Há a possibilidade de envio de uma única ORF ou várias em uma única submissão. Para esta última, é necessário especificar-se formatos para os dados, de modo a poder-se identificar e separar cada ORF presente. Neste caso, usamos o formato FASTA, seguindo as seguintes regras:

1. O início de uma ORF é determinado pela presença do chamado "caractere de início" que, por padrão é o sinal de maior, '>'.

2. O texto presente após o caractere de início até o próximo caractere de fim de linha é considerado como o código de identificação da ORF.

3. O texto presente entre o primeiro caractere de fim de linha (encontrado após o caractere de início) e o próximo caractere de início será considerado como a seqüência da ORF.

Exemplo de listagem de ORFs usando o formato FASTA:

```
>ORF1-gl-c-107-754
MGARLSKSYGNTMPVFCTREKLKKYVFSIVTDSRAPGEPKEAVGSPVFQLYQAFAGVE
ECSMFAQAFVVAFTTSESKRQDAKALGADEVVVSRDEESMAAHVKSFDLILNTVAASH
SLDPFLTLLKRDGTLTLVGAPATPHPSPQIFNLIFKRRSIAGSLIGGIAETQEMLDFC
AENGIVADIELIRADGINEAYERMMKGDVKYRFVIDNATLAA
>ORF2-gl-u-780-890
MLLNKRASGNPKKLPTRKYEIKKLCQITPAPLDIAGI
>ORF3-gl-u-880-972
MPEYESRSQPGTLIAIIKKKIFRRPLKTGFL
```

Observação: Caso os dados de ORF entrados não se iniciem com um caractere de início, considerar-se-á que está sendo feita diretamente a entrada da seqüência de aminoácidos, sem o texto de identificação. Assim, todo o texto presente do início até o primeiro caractere de início, se algum, é interpretado como sendo a seqüência da ORF.

Processamento e envio de Resultados

Os dados são enviados e devolvidos segundo a interface de comunicação CGI (Lemay, 1999; Herrman, 1997), *Common Gateway Interface*. Segundo a especificação desta interface, os dados preenchidos pelo usuário no formulário serão enviados pela Internet ao servidor que hospeda a página do *site*. Este servidor, então, executa o programa (desenvolvido neste projeto em linguagem Perl) responsável pela ferramenta enviando a ele os dados recebidos pela rede. O programa, segundo o algoritmo (ver seções abaixo) da ferramenta em questão, trabalha os tais dados e gera, internamente os resultados. Estes são então devolvidos ao servidor que os processa e envia de volta, pela Internet ao usuário.

Ao final do processo de submissão, o usuário recebe uma nova página HTML com os resultados dos cálculos, podendo visualizá-la imediatamente após seu recebimento.

Estimativas de pI e MW

Os valores de pI e MW das proteínas determinam sua localização nos géis 2D, conforme apresentado na Seção 1.1.1. No entanto, a partir das seqüências de aminoácidos das ORFs, estes valores podem ser calculados e, assim, servir de base para a elaboração de um mapa 2D teórico.

Como o pI é o pH onde a carga líquida da proteína é nula (Seção 1.1.1), o primeiro passo para seu cálculo é a obtenção de uma expressão para a carga de uma proteína. Pela teoria dos equilíbrios múltiplos[1], a carga líquida pode ser expressa pela Equação 2.1:

$$C_l = \sum_i \frac{Nb_i \cdot 10^{-pH}}{10^{-pH} + 10^{-pKb_i}} + \sum_i \frac{Na_i \cdot 10^{-pH}}{10^{-pH} + 10^{-pKa_i}} - Na_i \tag{2.1}$$

Na Equação 2.1, Nb_i é a quantidade de cada aminoácido básico; Na_i, a quantidade de cada aminoácido ácido; pKb, constante pK para aminoácidos básicos; $pKa.$ constante pK para aminoácidos ácidos; pH, variável independente da relação, indicando o pH que gera a carga líquida C_l. Esta equação depende então da composição de aminoácidos da seqüência (quantidade presente de cada um dos 20 aminoácidos) e dos valores de pK padrão dos aminoácidos, conforme apresentados na Tabela 2.1.

Tabela 2.1: Valores de pK para aminoácidos ionizáveis e terminais de seqüência. Para os demais aminoácidos têm-se $pK = 0$.

#	Aminoácido	pK
1	COOH-Terminal	3,55
2	NH3-Terminal	7,50
3	Arg (R)	12,00
4	Asp (D)	4,45
5	Cys (C)	9,00
6	Glu (E)	4,05
7	His (H)	6,25
8	Lys (K)	10,20
9	Tyr (Y)	10,00

A raiz da Equação 2.1 é, então, o pH onde a carga líquida da proteína é nula, ou seja, o pI. Esta raiz pode ser obtida por métodos numéricos (Kahaner *et al.*, 1989).

A relação 2.1 apresenta segunda derivada (para a variável pH) com sinal não constante no domínio de estudo (pH de 0 até 14), o que inviabiliza o método de Newton (Kahaner *et al.*, 1989; Scraton e Arnold, 1986) para se localizar a raiz. Em substituição, é utilizado o "Método das partições" (Kahaner *et al.*, 1989) do seguinte modo:

1. Testa-se o pH médio entre extremos da faixa de pH (inicialmente $pH = 0$ e $pH = 14$).

2. Sendo 0 o valor da equação 2.1 no pH testado, localizou-se o pI (nessa situação $pI = pH$) e termina-se o processo.

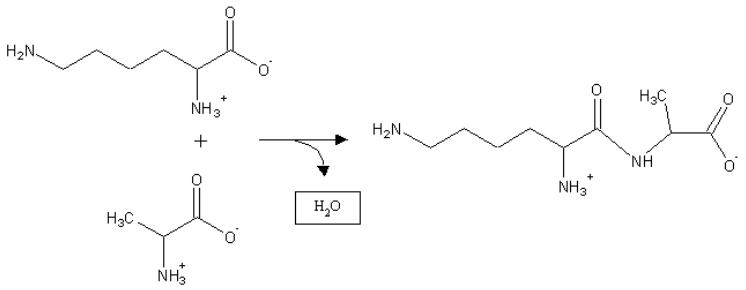
3. Sendo o valor da equação 2.1 maior que "0", muda-se o extremo inferior da faixa de pH para o pH testado e mantém-se o limite superior, modificando a faixa de busca, e inicia-se novamente o passo 1. Sendo o valor menor que "0", muda-se o

[1] http://www.up.univ-mrs.fr/~wabim, Mar/2005.

extremo superior da faixa de *pH* para o valor de *pH* testado e mantém-se o limite inferior, o que leva a uma nova faixa de *pH* e inicia-se o passo 1 novamente. A escolha de qual extremo da faixa é modificado se baseia no fato de que a equação 2.1 ser decrescente para a variável *pH*.

4. Também se finaliza o processo, localizando *pI* como o último *pH* testado quando, entre dois testes seguidos obtém-se o mesmo *pH* médio testado. Obtém-se seguidamente o mesmo valor médio quando os limites da faixa de *pH* sendo testada estão próximos o suficiente do zero para que o seu meio extrapole a precisão do computador (10 casas decimais).

Quanto à estimativa de MW, ela é baseada na soma algébrica da massa de cada aminoácido da seqüência, com a ressalva de que, na ligação química que dá origem à ligação peptídica que forma as proteínas, é perdida uma molécula de água. Isto é representado na equação 2.2 — formação de uma ligação peptídica entre os aminoácidos alanina e lisina. Levando este fator em conta, obtemos a equação 2.3 como forma de estimativa de MW.



(2.2)

$$MW = MW_{H_2O} + \sum_{i=1}^{20} n_i \cdot MW_i - MW_{H_2O} \tag{2.3}$$

Na Equação 2.3, temos que MW_{H_2O} é o peso molecular de uma molécula de água; i, o índice para cada um dos 20 aminoácidos existentes; N_i, o número de aminoácidos do tipo i existente na seqüência sendo considerada; MW_i, o peso molecular para aminoácido tipo i. Esta equação está em função da quantidade de cada aminoácido presente numa seqüência (n_i) e do peso molecular de cada aminoácido (MW_i), segundo a tabela 2.2.

Construção de Mapa 2D teórico

Os dados gerados pelas ferramentas de cálculo de pI e MW serão a base para a ferramenta de criação do mapa 2D teórico. Nesta, será montado, para um conjunto de valores de pI e MW, um gráfico (imagem) onde um dos eixos indicam os valores de pI e outro, os valores de MW. Cada ORF será representada por um ponto neste gráfico, sendo que tal ponto é dado pelo par ordenado de seu pI e seu MW.

Tabela 2.2: Valores de MW para aminoácidos

#	Aminoácido	Peso Molecular	#	Aminoácido	Peso Molecular
1	Ala	71,08	11	Pro	97,12
2	Gly	57,06	12	Val	99,14
3	Met	131,20	13	Glu	128,12
4	Ser	87,08	14	Lys	129,18
5	Cys	103,15	15	Gln	128,14
6	His	137,15	16	Trp	186,21
7	Asn	114,11	17	Phe	147,18
8	Thr	101,01	18	Leu	113,16
9	Asp	114,09	19	Arg	157,19
10	Ile	113,16	20	Tyr	163,18

Como resultado do processo de criação do mapa teórico, o programa produzirá uma imagem em formato GIF ou JPEG, que são formatos padrão de imagens na Internet (Lemay, 1999). Esta imagem, formada pelo gráfico do mapa 2D teórico, será o resultado a ser enviado pelo usuário.

Forma de análise dos resultados

Os dados obtidos pelas ferramentas de análise serão comparados com valores obtidos com ferramentas similares disponíveis na Internet. Porém serão utilizados, como fonte de dados de teste, outros genomas já estudados e não o da *Xylella fastidiosa*.

O uso da ferramenta de comparação de Gel Virtual com Real também pode, dentro de certa margem de restrição, ser usada como forma de validação da ferramenta de cálculo teórico de pI/MW. Pois a estimativa de pI/MW, apesar de teóricos, devem apresentar proximidade com valores que venham a ser obtidos experimentalmente.

2.1.2 Resultados

Algoritmos para os cálculos descritos acima foram implementados em *scripts* na linguagem Perl, utilizando tecnologia CGI para comunicar resultados de execução via Internet. E, com o desenvolvimento de páginas *web* para submissão de dados aos programas, estas ferramentas foram disponibilizadas via Internet[2]. A página de submissão disponibiliza campos para a entrada de informações às ferramentas desenvolvidas. Esses campos são:

Arquivo de ORFs deve-se informar um arquivo em formato FASTA com seqüências de ORFs a serem analisadas. Alternativamente à esse campo, há um campo para digitação do conteúdo do arquivo, onde não é necessário informar o nome de um arquivo em disco, mas pode-se entrar diretamente o conteúdo das ORFs.

[2]Ver em http://proteome.ib.unicamp.br/tools/pimw/index.html, Out/2006.

Opções de saída pode-se especificar os dados que se deseja obter como resultado da submissão:

- Título de cada ORF
- Seqüência de aminoácidos
- Resumo da ocorrência de aminoácidos
- Mapa 2D teórico

Valores de pK pode-se fazer uso de valores de pK padrão do *script*, que são os da tabela 2.1, ou especificar-se diretamente outros valores a serem usados nos cálculos.

O resultado da submissão é devolvido, segundo o padrão CGI, na forma de uma nova página HTML com o conteúdo especificado no formulário de submissão. A Figura 2.1

ORF:

XF0002 (XF-03E01-GL09)

Sequence:

MRFRLQRETFLKPLAHVVNVVERRQTRSIL ANLLIKVNEDQLSLTGTDLEVEMISKTIIE
DAESGEITIPARKIYEIVRALPDSSQLSVY QSDDKITLQAGRSRFTLATLPANDFPSIDK
IEVTERIIIPEVLLKELIERTAFAMAQQDV RYYLNGLLFDLRDTKLRCVATDGHRLALCE
TELEQAKDLKRQIILPRKGVMELQRLLEGS DRQIELEIARNHIRMKSFDVTFTSKLIDGS
FPDYEGVIPIGADREVKVAREVLRDALQRA AILSNEKYRGVRIEVSPGQLKINAHNPEQE
EAQEEIEAQTIVDGLAIGFNVNYLLDALSS LRGDFVNIQLRDSNSSALIRESNSEKSLQV VMPLRL

MW:	pI:
41549.68	5.36

Amino-acid composition

Ala (A)	26	7.1%	Met (M)	6	1.6%
Cys (C)	2	0.5%	Asn (N)	14	3.8%
Asp (D)	23	6.3%	Pro (P)	12	3.3%
Glu (E)	34	9.3%	Gln (Q)	19	5.2%
Phe (F)	11	3.0%	Arg (R)	32	8.7%
Gly (G)	15	4.1%	Ser (S)	24	6.6%
His (H)	4	1.1%	Thr (T)	17	4.6%
Ile (I)	32	8.7%	Val (V)	25	6.8%
Lys (K)	17	4.6%	Trp (W)	0	0.0%
Leu (L)	46	12.6%	Tyr (Y)	7	1.9%

Total: 366

Figura 2.1: Página de resultado da ferramenta para cálculo de pI e MW. É mostrada tabela com informações sobre as seqüências submetidas, incluindo pI, MW, seqüência de aminoácidos e composição de aminoácidos.

Ao contrário de outras ferramentas disponíveis para estimativa de pI e MW, a ferramenta desenvolvida têm o recurso de trabalhar sob um conjunto de várias ORFs que devem ser submetidas sob o formato FASTA. Este recurso é imprescindível para se poder implementar uma ferramenta de mapa 2D teórico, como fica evidente adiante.

Implementados os algoritmos de estimativa de pI/MW, pôde-se desenvolver a ferramenta para a construção de mapa 2D teórico: dados pares ordenados (pI, MW) constrói-se o gráfico representando o gel 2D com eixo horizontal numa graduação de pH e eixo vertical numa graduação de MW plotando-se para cada par um ponto.

O gel 2D teórico é gerado numa figura de formato GIF que é devolvida como resultado da submissão ao *site*, conforme exemplifica a Figura 2.2. Para se gerar a figura em

formato GIF foi utilizada a ferramenta "FLY" [3] que constrói um arquivo GIF a partir de saída gerada na linguagem Perl.

Esta figura apresenta uma escala padrão, entretanto, para efeito de comparações, o mapa 2D deve apresentar uma mesma escala que géis de referência obtidos experimentalmente. Assim, apesar de ainda não disponível no *site*, foram realizadas modificações nos procedimentos iniciais de plotagem de pontos no mapa 2D de modo a se adaptar a escala de plotagem. Para isso, pontos de referência, com suas coordenadas sobre a imagem original e seu correspondente valor de pI ou MW, são obtidos no gel experimental e estes são usados como referência de localização de pI ou MW sobre a imagem 2D teórica a ser plotada. Esta modificação permitiu a comparação direta de géis de referência com géis 2D teóricos.

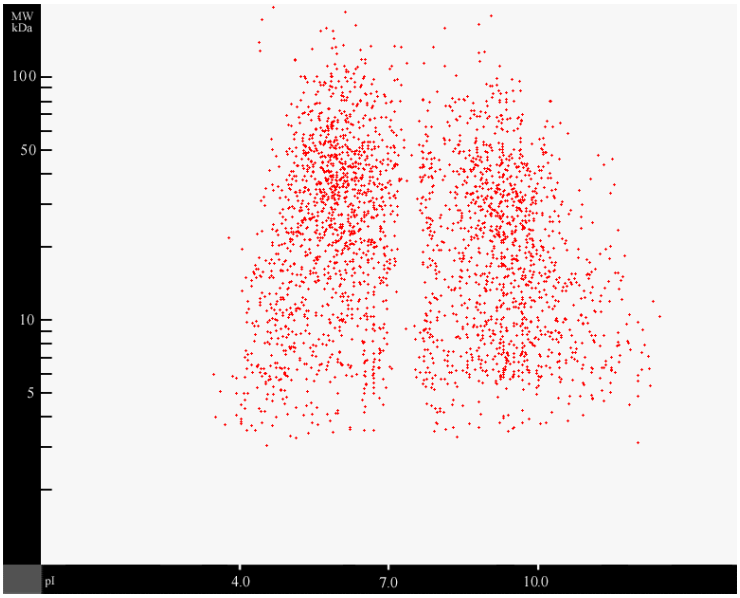


Figura 2.2: Mapa 2D teórico criado pela ferramenta de predição de pI e MW. Cada ponto representa um spot para uma proteína cuja seqüência foi submetida.

2.1.3 Testes

Implementadas as ferramentas para cálculo de pI/MW, pôde-se iniciar testes sobre a proximidade de suas estimativas com valores experimentais. Inicialmente foram realizados testes comparativos com outras ferramentas para estimativas de pI e MW já existentes e, posteriormente, teste com conjuntos de proteínas já identificadas e suas seqüências.

[3] http://martin.gleeson.com/fly/, Out/2006.

Testes com outros *sites*

Estes testes permitiram ajustes em erros de implementação e obter experiência de fontes que já vêm utilizando ferramentas similares.

Para os testes são utilizados diferentes conjuntos de valores de pK, este que é um parâmetros para a equação 2.1.

Tabela 2.3: Valores de pI calculados (Lehninger, Tanford, Sillero, Citi2, Bull e Bjelqvist) e valores obtidos de *sites* externos (Abi e Expasy)

Teste	Lehninger	Tanford	Sillero	Citi2	Bull	Bjelqvist	Abi	Expasy
ORF 1	5,33	5,7	5,6	5,8	5,45	5,51	5,26	5,47
ORF 2	10,59	10,44	10,47	10,44	10,93	10,3	10,5	10,14
ORF 3	10,99	10,81	10,84	11,08	11,33	10,69	10,89	10,55
ORF 4	5,25	5,57	5,55	5,86	5,51	5,51	5,12	5,41
ORF 5	9,95	9,6	9,71	9,86	10,23	9,56	9,95	9,4
ORF 6	8,8	7,85	8,52	8,94	8,42	8,18	9,01	7,93
ORF 7	10,1	9,92	9,92	9,87	10,31	9,72	10,08	9,52
ORF 8	5	5,44	5,24	5,47	5,07	5,24	4,99	5,17
ORF 9	11,37	11,1	11,1	11,56	11,55	11	11,36	10,9
ORF 10	7,93	8,16	8,32	8,21	7,97	8,21	7,98	8,13
ORF 11	5,73	5,96	6,09	6,42	6,12	5,89	5,61	5,77
ORF 12	11,38	11,12	11,12	11,19	11,57	11,02	11,38	10,93
ORF 13	10,76	10,63	10,65	10,59	11,09	10,47	10,7	10,29
ORF 14	11,12	10,92	10,93	10,98	11,4	10,81	11,16	10,68
ORF 15	9,74	9,47	9,56	9,49	10,04	9,39	9,73	9,22
ORF 16	7,65	7,76	8,04	8,02	7,76	7,87	7,67	7,69
ORF 17	11,37	11,1	11,1	11,56	11,55	11	11,36	10,9
ORF 18	4,67	5,17	4,87	5,15	4,71	4,95	4,61	4,86
ORF 19	7,78	7,8	8,13	8,12	7,82	7,93	7,83	7,74
ORF 20	4,2	4,73	4,4	4,69	4,26	4,5	4,17	4,42
ORF 21	10,69	10,55	10,57	10,53	11,03	10,39	10,61	10,22
ORF 22	4,52	4,93	4,77	4,9	4,58	4,58	4,4	4,72
ORF 23	6,87	6,6	7,19	7,36	7,04	6,88	6,86	6,55
ORF 24	3,94	4,49	4,13	4,46	4	4,27	3,93	4,15
ORF 25	2,89	3,64	3,37	3,63	3,39	3,37	3,06	3,5

Particularmente, a estimativa realizada pelo *site* ExPASy é preferida pelos pesquisadores do Projeto do Proteoma da *Xylella Fastidiosa*, o levou a se definir os valores de pK que mais se aproximassem das estimativas desse *site* como padrão.

Entretanto os testes para diferentes valores de pK, conforme a Tabela 2.3 inclusive os indicados por bibliografia do *site*, apresentavam sempre divergências. Isso levou à busca em erros no *script*, o que não foi confirmado, e propostas de modelos diferentes para a estimativa e re-análise do modelo.

Segundo Patrickios e Yasaki (1995), http://www.up.univ-mrs.fr/∼wabim (Set/2005) e www.auths.edu/local/gcghelp/isoelectric.html (Mar/2005), que expõem métodos de es-

timativa de pI que concordam com a forma de estimativa aqui utilizada, a suspeita por diferenças nos valores de pK se mostravam como a mais provável fonte da divergência.

Nos cálculos, conforme deixa evidente a tabela 2.1, utilizava-se sempre os mesmos valores de pK para os terminais Amina e Carboxila (N-terminal e C-terminal, respectivamente), entretanto estes terminais vêm de aminoácidos diferentes, de proteína para proteína, conforme o aminoácido final ou inicial sejam diferentes. Isso levou a se propor que valores de pK diferentes para N-terminal e C-terminal deveriam ser usados conforme o aminoácido iniciador ou finalizador da seqüência. Posterior contato com a equipe do *site* ExPASy mostrou que este método era utilizado em seus cálculos confirmando a proposta.

Com essa modificação a tabela 2.1 ganhou duas novas colunas que indicam os diferentes pK de N-terminal e C-terminal, conforme a tabela 2.4.

Tabela 2.4: Valores de pK, inclusive para terminais, para os 20 aminoácidos.

Aminoácidos		R[a]	C-Terminal[b]	N-Terminal[c]
A	Ala	0,0	3,55	7,59
C	Cys	9,0	3,55	7,5
D	Asp	4,05	4,55	7,5
E	Glu	4,45	4,75	7,7
F	Phe	0,0	3,55	7,5
G	Gly	0,0	3,55	7,5
H	His	5,98	3,55	7,5
I	Ile	0,0	3,55	7,5
K	Lys	10,0	3,55	7,5
L	Leu	0,0	3,55	7,5
M	Met	0,0	3,55	7,0
N	Asn	0,0	3,55	7,5
P	Pro	0,0	3,55	8,36
Q	Gln	0,0	3,55	7,5
R	Arg	12,0	3,55	7,5
S	Ser	0,0	3,55	6,93
T	Thr	0,0	3,55	6,82
V	Val	0,0	3,55	7,44
W	Trp	0,0	3,55	7,5
Y	Tyr	10,0	3,55	7,5

[a] *pK* do resíduo do aminoácido
[b] *pK* usado para a porção C-Terminal, conforme qual aminoácido finaliza a seqüência.
[c] *pK* usado para a porção N-Terminal, conforme qual aminoácido inicia a seqüência.

Os valores estimados segundo este novo procedimento e com um conjunto de pKs segundo a tabela 2.4 passaram a diferir por 0,01 no máximo aos calculados pelo ExPASy, sendo que este desvio deve-se provavelmente ao uso de diferentes métodos de arredondamento e restringe-se a última casa decimal apresentada.

As estimativas de MW apresentaram desvios percentuais numa média de 0.2% quando comparados com estimativas do *site* ExPASy, estando portanto muito próximos.

Testes sobre dados experimentais

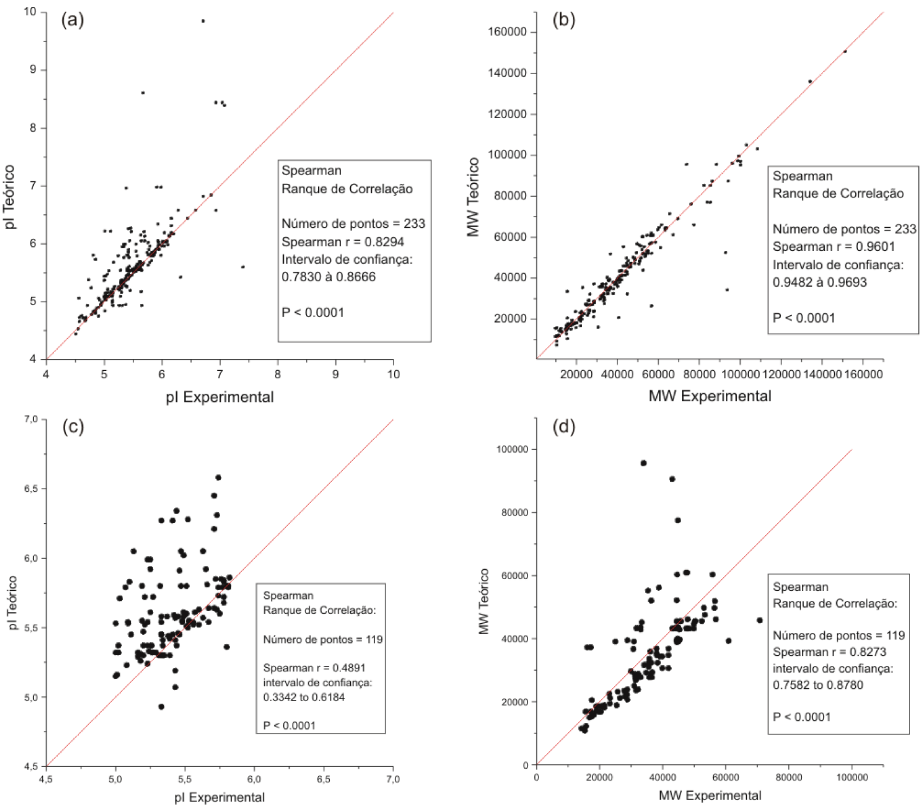


Figura 2.3: Comparação entre pI e MW experimentais e teóricos de proteínas da bactéria *Escherichia coli*. Nos painéis a (pI) e b (MW) as proteínas foram obtidas de um gel 2D de referência de pH de 4,5 à 8. Já em c (pI) e d (MW) usou-se um gel com faixa de pH de 5 à 6. A proximidade dos pontos com a linha de 45º é uma medida da concordância entre valores teóricos e experimentais.

Foi obtido uma listagem com 233 proteínas do proteoma da bactéria *Escherichia coli* [45], tal listagem, entre outras informações, contém a seqüência de aminoácidos e valores experimentais de pI e MW para cada proteína. Com tais informações pôde-se obter correspondes valores teóricos de pI e MW, além do gel 2D teórico. Estes dados teóricos foram comparados com os dados experimentais, analises das comparações podem ser acompanhadas na Figura 2.3.

Os painéis *a* e *b* da Figura 2.3 mostram a boa correlação entre experimentos e previsões baseadas na teoria usada para estimativa de pI e MW para um gel 2D numa faixa de pH de 4,5 à 8 da *Escherichia coli*. No painel *a*, 139 dos 233 pontos apresentam

[4] http://expasy.cbr.nrc.ca/cgi-bin/get-ch2d-table.pl, Mar/2005.
[5] http://expasy.cbr.nrc.ca/sprot/, Mar/2005.

um desvio absoluto em relação ao valor experimental menor que 0,1; de modo geral, este painel apresenta um desvio absoluto médio de 0,22, comparável com resultados encontrados em testes semelhantes publicados (Patrickios e Yasaki, 1995). O painel *b* apresenta a comparação entre teoria e experimento para valores de MW com um desvio relativo médio de 9,2%. Também foram realizados testes para um gel de referência da mesma bactéria para uma faixa de pH de 5 à 6 (painéis c e d da Figura 2.3). O painel *c* apresenta um desvio médio absoluto de 0,24 na escala de pH e o painel *d* um desvio relativo médio de 20%.

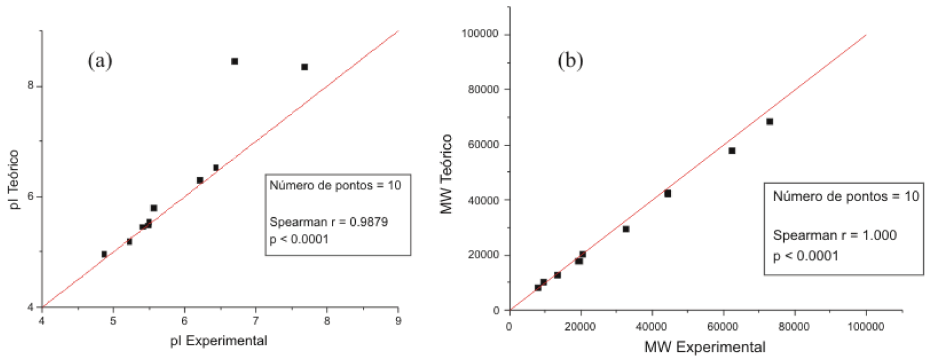


Figura 2.4: Comparação entre pI (a) e MW (b) experimentais e teóricos de 10 proteínas já identificadas da bactéria *Xylella fastidiosa*. A proximidade dos pontos com a linha de 45º é uma medida da concordância entre valores teóricos e experimentais.

Comparações com dados experimentais e dados extraídos do genoma da *Xylella fastidiosa* também foram realizados. Há um pequeno número de proteínas disponíveis já identificadas, assim foram utilizadas 10 proteínas para verificações das estimativas de pI e MW, os primeiros resultados podem ser observados na Figura 2.4. Há uma boa correlação entre dados experimentais e teóricos sendo que, para as comparações de pI, houve um desvio absoluto médio de 0,3 na escala de pH e, para MW, um desvio relativo médio de 5,8%, valores comparáveis aos obtidos anteriormente com dados da bactéria *Escherichia coli*.

Além das comparações para pI e MW, foi construído o mapa 2D teórico para a *Xylella fastidiosa*. Esse mapa é comparado com o obtido pela plotagem de pares (pI, MW) de spots em sua maioria não identificados obtidos da análise da imagem do gel de referência (dado experimental) da mesma bactéria, o resultado é observado na Figura 2.5. O resultado é semelhante a testes semelhantes publicados (Urquhart *et al.*, 1998), com a presença de duas regiões de concentração de pontos teóricos, uma faixa de pH de entre 4 e 7 e outra de 7,5 à 10, sendo que os pontos experimentais se concentram na primeira faixa. Notadamente, os organismos vivos não expressam todas proteínas previsíveis pelo seu genoma e há dificuldades para observação nos géis 2D de todas proteínas expressas (Urquhart *et al.*, 1998), o que vem a explicar a quantidade de pontos teóricos expressivamente superior à de pontos experimentais observados na Figura 2.5.

A comparação entre géis 2D teóricos e experimentais pode ser facilitada se, diferentemente do realizado na Figura 2.5 onde houve uma plotagem de pares (pI, MW)

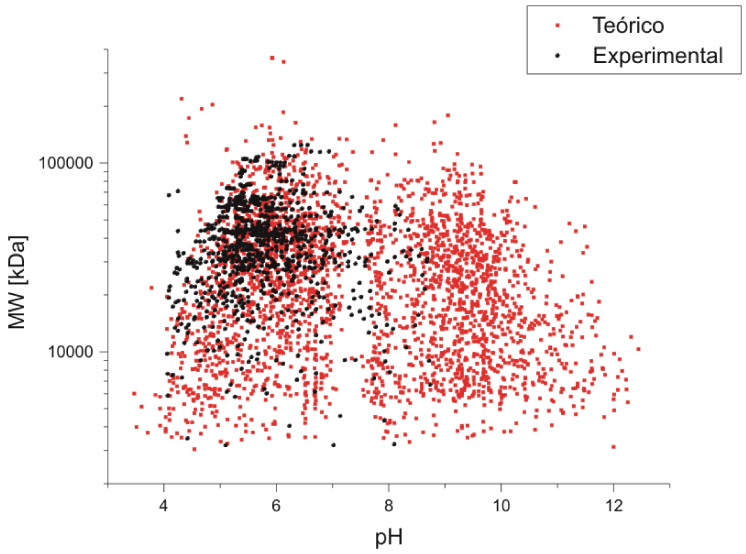


Figura 2.5: Mapa 2D teórico comparado com plotagem de pontos (pI, MW) obtidos da análise de imagem de gel de referência para a bactéria *Xylella fastidiosa*.

experimentais, for feita a plotagem direta dos pontos teóricos sobre a imagem do gel de referência. Assim a ferramenta inicial de construção de mapa 2D teórico foi modificado de modo a permitir ajustes de escala em sua plotagem, característica necessária visto que esta é realizada no gel de referência. Para este ajuste de escala, são fixados inicialmente as posições (em pixeis) de referência sobre a imagem final por onde deve-se constar determinado valor de pI (ou MW), essas posições de referência são obtidas do gel de referência em momento anterior a execução do *script* para a construção do mapa. Na proposta implementada até o momento, a localização de cada valor de pI (ou MW) encontra-se numa extrapolação linear entre as posições de pontos de referência imediatamente superior e inferior a cada valor de pI (ou MW). Os resultados para o gel de referência da *Xylella fastidiosa* podem ser observados na Figura 2.6.

Conforme discutido no projeto inicial, a identificação de proteínas envolve a combinação de várias técnicas de análise tanto das proteínas em si, quanto de dados teóricos de ORFs. Dentre as ferramentas de análises teóricas foram desenvolvidas as de estimativa de valores de ponto isoelétrico e peso molecular e criação de mapa 2D teórico.

2.2 Estatísticas de N-terminais

Um recurso que complementa as estimativas teóricas de pI/MW (aliadas aos géis 2D experimentais) no processo de identificação de proteínas é o seqüenciamento N-terminal, ver Seção 1.1.2.

Como recurso de apoio a esta técnica, foi desenvolvido e implementado, também em linguagem Perl para funcionamento via CGI, um algoritmo para geração de listas

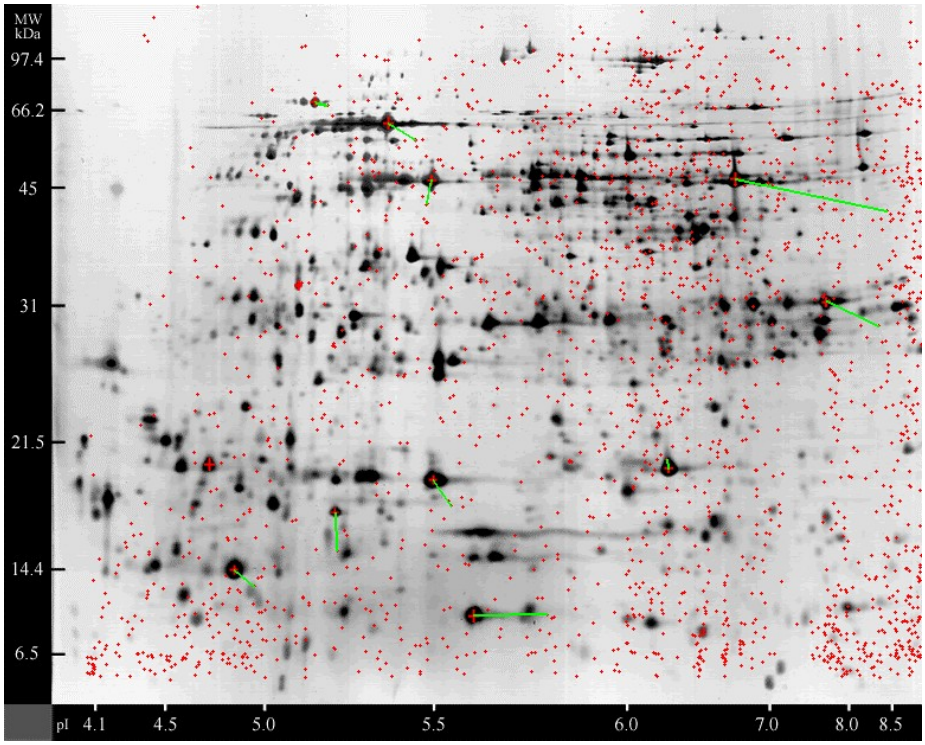
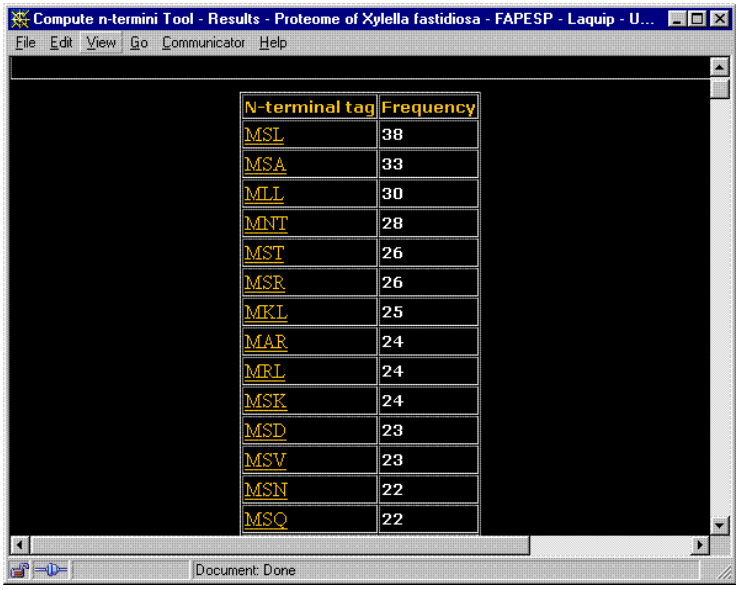


Figura 2.6: Mapa 2D teórico, da *Xylella fastidiosa*, criado pela ferramenta desenvolvida. Os pontos teóricos (vermelhos) foram plotados contra a figura do gel de referência dessa bactéria (pontos pretos). Spots já identificados do gel de referência aparecem com '+' vermelho. Em adição ao gel de referência e ao teórico plotado, foram desenhadas barras vermelhas ligando esses spots identificados com seus respectivos pontos teóricos. O gel de referência sofre tratamentos que ajustam posições de spots não identificados com base em já identificados, resultando numa imagem com escalas não padrão (observa-se, por exemplo, na escala de pH, que a distância visual entre o pH 4,5 e 5,0 é menor que entre 5,5 e 6,0). Assim, a plotagem dos pontos teóricos, pela ferramenta de mapa 2D teórico, é feita considerando-se as escalas do gel de referência, sem a qual a comparação visual entre ambos géis seria inviabilizada.

de N-terminais e acesso às ORFs correspondentes a cada seqüência N-terminal. O dado de entrada principal para o *script* desenvolvido é um arquivo em formato FASTA com as diferentes ORFs a serem analisadas. Como dado de entrada secundário, mas não facultativo, tem-se o tamanho da seqüência ou uma seqüência de aminoácidos para comparação.

Caso seja informado o tamanho da seqüência, o *script* seleciona, de cada ORF presente no arquivo FASTA submetido, a quantidade de letras (do início da seqüência de aminoácidos) correspondente ao tamanho informado, formando uma subseqüência. Para cada subseqüência encontrada soma-se sua ocorrência numa tabela de ocorrências. De modo que, ao final de um processamento de todo o arquivo submetido, tem-se registrada a quantidade de ORFs que apresentam a mesma subseqüência inicial (N-terminal), para cada subseqüência diferente presente no arquivo.



N-terminal tag	Frequency
MSL	38
MSA	33
MLL	30
MNT	28
MST	26
MSR	26
MKL	25
MAR	24
MRL	24
MSK	24
MSD	23
MSV	23
MSN	22
MSQ	22

Figura 2.7: Página com a tabela de resultado para a listagem N-terminal gerada pelo Primeiro *script* desenvolvido para pesquisa de N-terminais. Neste exemplo, foi submetido o arquivo FASTA com as 2.847 ORFs do genoma da *Xylella fastidiosa*, na figura vê-se o topo da listagem que contém no total 376 seqüências N-terminais distintas de 3 aminoácidos.

Alternativamente, se for informada uma seqüência de aminoácidos para comparação, somente são utilizadas, do arquivo submetido, aquelas ORFs cuja seqüência de aminoácidos inicial (o N-terminal) for idêntico à seqüência informada. Opcionalmente, na seqüência informada pode ser incluído, além das letras correspondentes a cada um dos 20 aminoácidos, um (ou mais) caracter especial, no caso "X". O caracter especial indica uma posição na seqüência informada, onde pode haver qualquer aminoácido. Por exemplo, uma seqüência MRFX, seria identificada com MRFA, MRFG, MRFM, MRFS, etc. Dado as ORFs selecionadas pela seqüência de comparação, o pro-

cesso de contagem se dá como descrito anteriormente.

O recurso de se informar uma seqüência para comparação possibilita a otimização do sequênciamento N-terminal, facilitando a identificação da ORF ainda durante o sequênciamento de aminoácidos. Assim, com a ajuda do software, é possível verificar automaticamente as ocorrências da seqüência geradas. De outro modo, não se informando tal seqüência de comparação, o *script* trabalha sobre todas as ORFs, devolvendo um resultado que permite se ter uma visão geral das várias seqüências N-terminais possíveis do genoma em estudo.

Para diponibilizar o acesso ao *script* de listagem N-terminal, foi desenvolvida uma página de submissão com campos para se informar o arquivo FASTA a ser processado, o tamanho da seqüência e a seqüência de comparação. Esta página, que pode ser acessada via http://proteome.ibi.unicamp.br/tools/nterm-work.html, segundo as especificações CGI, passa as informações destes campos ao *script* desenvolvido. E, como resposta a solicitação da página, ele retorna uma nova página HTML com a tabela de freqüência de ocorrência de N-terminais calculada.

A partir desta primeira página de resultados (exemplo: Figura 2.7), vê-se cada seqüência N-terminal encontrada e sua freqüência. A partir daí, foi desenvolvido novo programa para permitir visualizar a listagem das ORFs correspondentes a cada entrada obtida na página de resultados. Para isso adotou-se os seguintes passos:

1. O *script* já desenvolvido (1º *script*) foi modificado para que salve todo o arquivo FASTA a ele submetido.

2. A primeira página de resultados (Figura 2.7), é retornada com o acréscimo de informações sobre a localização do arquivo já salvo (item 1). Este acréscimo é feito pelo 1º *script* ao gerar esta página.

3. Desenvolve-se um novo *script* (2º *script*) que, recebendo uma série de letras correspondentes aos aminoácidos de uma seqüência N-terminal, mais a indicação da localização de um arquivo FASTA armazenado, seleciona todas as ORFs do arquivo indicado que possuam a seqüência N-terminal idêntica a série de letras (aminoácidos) recebida.

4. Ao se selecionar, na primeira página, uma determinada seqüência N-terminal, esta, juntamente com as informações presentes na página sobre a localização do arquivo, é passada ao 2º *script*, que assim retorna uma nova página com a listagem das ORFs que possuem o mesmo N-terminal selecionado.

Implementandos estes passos, atingiu-se o esperado: obter, automaticamente, todas as ORFs, do arquivo FASTA original, de uma mesma seqüência N-terminal de interesse. Um exemplo de resultado do 2º *script* desenvolvido é observado na Figura 2.8.

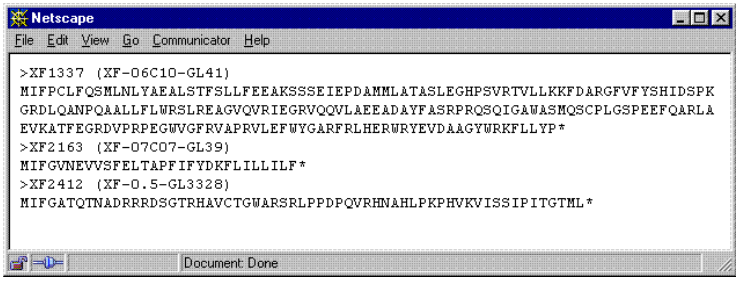


Figura 2.8: ORFs selecionadas pelo 2º *script* desenvolvido para listagens N-terminais. Ao *script* foi submetido o arquivo FASTA da *Xylella fastidiosa* e foi solicitada a seleção das ORFs com N-terminal "MIF", como mostra a figura, há 3 ORFs no genoma desta bactéria com tal N-terminal.

Capítulo 3

Digestão teórica e identificação de peptídeos

3.1 Métodos

Esta seção envolve duas fases de atividades, primeiramente uma de avaliação da estratégia de identificação de peptídeos e outra de desenvolvimento de ferramenta de busca.

Com base nelas, os resultados teóricos e estatísticos obtidos de verificação da seletividade do método serão avaliados. Também é verificada, em testes sobre dados experimentais e simulações, a capacidade da ferramenta de busca localizar uma proteína sendo estudada.

3.1.1 Visualização de peptídeos

Buscando-se obter uma visão global sobre distribuição dos valores de pI e MW dos peptídeos, foi também proposta a criação gráficos de duas dimensões com eixos de pI e MW, o que equivaleria a uma eletroforese teórica bidimensional dos peptídeos.

A ferramenta de mapa teórico já desenvolvido (ver Capítulo 2), assim, usada para a plotagem destes gráficos. Sendo que, como ela exige a entrada de dados em formato FASTA, foi desenvolvido um novo *script* para a conversão dos resultados da digestão teórica para este formato.

3.1.2 Avaliação teórica da estratégia

Nesta fase, foi desenvolvida uma ferramenta computacional (programa de computador) para realização da digestão teórica de uma lista de seqüências de aminoácidos e que faz a seleção de peptídeos que contenham cisteína, correspondendo a presença do caracter "C" na seqüência. Além disso, as ferramentas de cálculo de pI e MW (Capítulo 2) foram reutilizadas na confecção de filtros (novos programas) para seqüências de aminoácidos dentro de uma faixa de valores a ser fornecida como parâmetro. Isso permite a seleção de proteínas e peptídeos por faixa de MW e pI, respectivamente. Com essas

ferramentas, as seleções correspondentes aos passos 2, 3, 4 e 5, descritos na Seção 1.1.4 e 6 ficam cobertas.

Assim, concluindo a fase, as ferramentas são aplicadas, em conjunto, sobre as ORFs da bactéria *Xylella fastididosa* para obter estatísticas sobre seletividade do método e o tamanho da faixa de pI e MW, a serem utilizadas nos passos 2 e 5, suficiente para se conseguir identificação de proteínas.

3.1.3 Desenvolvimento de ferramenta de busca

Na fase final, foi desenvolvida uma ferramenta computacional, com disponibilização via Internet, para realização de busca de proteínas segundo os critérios da abordagem sugerida. Ela receberá, como parâmetros, os valores de:

- faixa de MW.
- faixa de pI.
- massa molecular de peptídeo.

Esses valores, em conjunto com os critérios de geração de peptídeos por digestão tríptica e a informação da presença de cisteína em cada peptídeo, foram então usados para selecionar proteínas. Assim, esta ferramenta é capaz de executar a fase *in silico* de um experimento biológico para identificação de proteínas dentro da estratégia discutida.

3.2 Resultados

3.2.1 Obtenção de peptídeos

Anteriormente ao processo de avaliação do método, é necessária a obtenção das próprias seqüências de peptídeos.

Ferramentas disponíveis na Internet, como o "Peptide Mass" do servidor ExPASy do Swiss Institute of Bioinformatics[1], fornecem meios para a realização de digestão teórica de seqüências de proteínas. Mas não são suficientes para nossas necessidades uma vez que trabalham sobre uma única proteína fornecida por um usuário[2] [3] e necessitamos da digestão de um conjunto inteiro de ORFs de um genoma — 2.830 ORFs no caso da *Xylella fastidiosa*. Também são necessárias algumas especificidades não encontradas em ferramentas já disponíveis, como o cálculo da massa de peptídeos e proteínas levando-se em conta a presença de moléculas de ICAT ligadas às cisteínas (o que causa um acréscimo na massa de cada cisteína de 442,225 Da, no caso da forma leve, e 450,275 Da, no caso da forma deutérica); também será necessária a estimativa teórica do pI de tais peptídeos, opção não encontrada nestas ferramentas.

[1] http://ca.expasy.org/tools/peptide-mass.html, Fev/2005.
[2] http://ca.expasy.org/tools/peptide-mass, Fev/2005.
[3] http://prospector.ucsf.edu/ucsfhtml4.0u/msdigest.htm, Fev/2005.

Deste modo, buscou-se desenvolver um *script*, ou programa, de computador para a realização da digestão segundo nossas necessidades, tendo, entretanto, as ferramentas citadas como referência.

O *script* desenvolvido realiza digestão teórica por tripsina, quebrando seqüências de aminoácidos após as ocorrências de lisinas (K) e argininas (R), com exceção dos casos onde há prolinas (P) após as lisinas e argininas[4]. Em experimentos reais, há a possibilidade de ocorrência de falhas de quebra (*missed cleaves*), por isso o programa considera a possibilidade de até uma falha de quebra na digestão, por local de quebra.

Ele também calcula a massa molecular e o ponto isoelétrico dos aminoácidos criados. Como a massa dos peptídeos deverá ser medida, em experimentos, por espectrômetro de massa de alta resolução, o cálculo de MW faz uso das massas monoisotópicas dos aminoácidos, segundo a Tabela 3.1.

Tabela 3.1: Massas moleculares, médias e monoisotópicas, dos 20 aminoácidos (http://ca.expasy.org/tools/findmod/findmod_masses.html, Jul/2005)

Aminoácido	Mono.	Média	Aminoácido	Mono.	Média
Alanina (A)	71,04	71,08	Leucina (L)	113,08	113,16
Arginina (R)	156,1	156,19	Lisina (K)	128,09	128,17
Asparagina (N)	114,04	114,1	Metionina (M)	131,04	131,19
Ác. Aspártico (D)	115,03	115,09	Fenilalanina (F)	147,07	147,18
Cisteína (C)	103,01	103,14	Prolina (P)	97,05	97,12
Ác. Glutâmico (E)	129,04	129,12	Serina (S)	87,03	87,08
Glutamina (Q)	128,06	128,13	Treonina (T)	101,05	101,11
Glicina (G)	57,02	57,05	Triptofano (W)	186,08	186,21
Histidina (H)	137,06	137,14	Tirosina (Y)	163,06	163,18
Isoleucina (I)	113,08	113,16	Valina (V)	99,07	99,13

Para os cálculos de MW, também foi acrescida, à massa das cisteínas, a massa da molécula de ICAT, segundo a Tabela 3.2. Na adição, não se deve considerar a massa total dele uma vez que, na ligação química, perde-se um iodo.

Tabela 3.2: Massas moleculares do reagente ICAT (Biosystems, 2001).

Tipo do ICAT	Média	Monoisotópica	Acrescenta-se à cisteína
ICAT leve (D0)	570,5 Da	570,1373 Da	442,2250 Da
ICAT pesado (D8)	578,5 Da	578,1875 Da	450,2752 Da

Além disso, deve-se observar que, como a ligação de ICAT inibe as propriedades de ionização dos radicais de cisteínas, elas não podem ser consideradas durante o cálculo de pI.

Um exemplo de resultados de obtenção de peptídeos pelo *script* está na Tabela 3.3, que é um trecho da digestão teórica da seguinte seqüência de aminoácidos:

[4] http://ca.expasy.org/tools/peptide-mass-doc.html, 02/2005.

```
MESWSRCLERLETEFPPEDVHTWLRPLQADQRGDSVVLYAPNPFIIELVEERYLGRLRELLSYFSGIREVVLAIG
SRPKTTELPVPVDTTGRLSSTVPFNGNLDTHYNFDNFVEGRSNQLARAAAWQAAQKPGDRTHNPLLLYGGTGLGK
THLMFAAGNVMRQVNPTYKVMYLRSEQFFSAMIRALQDKSMDQFKRQFHQIDALLIDDIQFFAGKDRTQEEFFHT
FNALFDGKQQIILTCDRYPREVNGLEPRLKSRLAWGLSVAIDPPDFETRAAIVLAKARERGATIPDEVAFLIAKK
MHSNVRDLEGALNTLVARANFTGRAVTIEFSQETLRDLLRAQQQTIGIPNIQKIVADYYGLQIKDLLSKRRTRSL
ARPRQLAMALAKELTEHSLPEIGDAFAGRDHTTVLHACRQIKLLMETETKLREDWDKLMRKFSE
```

Tabela 3.3: Trecho inicial do resultado da digestão teórica de uma ORF, considerando uma falha de quebra pela tripsina.

Início	Final	MW	Cisteínas	pI	Seqüência
1	6	794,34	0	5,75	MESWSR
1	10	1737,8	1	5,9	MESWSRCLER
7	10	961,47	1	6	CLER
7	32	3619,79	1	4,64	CLERLETEFPPEDVHTWLRPLQADQR
11	32	2676,33	0	4,5	LETEFPPEDVHTWLRPLQADQR
11	52	4917,5	0	4,25	LETEFPPEDVHTWLRPLQADQRGDSVVLY-APNPFIIELVEER
33	52	2259,18	0	4	GDSVVLYAPNPFIIELVEER
33	56	2748,45	0	4,41	GDSVVLYAPNPFIIELVEERYLGR
53	56	507,28	0	8,75	YLGR
53	58	776,47	0	10,84	YLGRLR
57	58	287,2	0	9,75	LR
57	68	1452,81	0	8,75	LRELLSYFSGIR
59	68	1183,62	0	6,1	ELLSYFSGIR
59	79	2333,31	0	8,69	ELLSYFSGIREVVLAIGSRPK
69	79	1167,7	0	8,85	EVVLAIGSRPK
69	92	2534,41	0	6,28	EVVLAIGSRPKTTELPVPVDTTGR
80	92	1384,72	0	4,37	TTELPVPVDTTGR
80	116	4108,98	0	4,43	TTELPVPVDTTGRLSSTVPFNGNLDTHYN-FDNFVEGR
93	116	2742,27	0	4,54	LSSTVPFNGNLDTHYNFDNFVEGR
93	122	3411,62	0	5,38	LSSTVPFNGNLDTHYNFDNFVEGRSNQLAR
117	122	687,37	0	9,47	SNQLAR
117	135	2038,05	0	10,83	SNQLARAAAWQAAQKPGDR

3.2.2 Formas de registro de peptídeos

Uma vez gerados os peptídeos, estes devem ser armazenados formando um registro dos peptídeos de todas as ORFs sendo estudadas. Aqui, usou-se as ORFs do genoma da bactéria *Xylella fastidiosa*. Mantendo-se registrados todos os peptídeos, economiza-se tempo sobre as subseqüentes análises que são feitas sobre estes dados. Caso contrário, a cada consulta ou busca selecionando-se peptídeos, seria necessário refazer a digestão, processo que durou em torno de 6 minutos nos testes realizados em nosso laboratório. Cabe ressaltar também que a forma como estes dados de peptídeos são armazenados podem ser determinantes no desempenho de consultas aos mesmos.

A primeira forma testada para armazenamento consistiu na criação, para cada ORF, de um arquivo com a listagem de seus peptídeos em formato padronizado e do tipo

texto. Desta forma, foram criados, como resultado da digestão, 2.830 arquivos nomeados de forma a identificar a ORF que o originou, num total de 149.946 peptídeos. Um exemplo do conteúdo destes arquivos está na listagem abaixo, que mostra os peptídeos da ORF XF0006 que contém cisteínas:

```
299- 306:  1194.8718 -  6.71 - VPHCSPAA
295- 306:  1594.0585 -  6.74 - DAGRVPHCSPAA
102- 116:  2099.3736 -  4.37 - CVVDALVAELPSQWR
102- 119:  2454.5956 -  6.07 - CVVDALVAELPSQWRQVK
 39-  61:  2727.7168 -  9.75 - LSFFLIATLSLGACSTATSPTGR
 96- 116:  2830.7451 -  6.07 - QNAYVRCVVDALVAELPSQWR
 38-  61:  2855.8118 - 11.00 - KLSFFLIATLSLGACSTATSPTGR
  4-  31:  3593.1278 -  4.00 - LLDFISLGCTNFYETLNAFEETLTALIK
  4-  33:  3863.2718 -  4.41 - LLDFISLGCTNFYETLNAFEETLTALIKNR
  1-  31:  3979.3378 -  4.41 - MVRLLDFISLGCTNFYETLNAFEETLTALIK
 39-  85:  5423.0886 -  6.76 - LSFFLIATLSLGACSTATSPTGRHQVFSGVS
                               QQQLNQLGEQAFVEIK
```

Esta estrutura de dados também conta com um índice de massas para as ORFs, o que agiliza a localização de proteínas com MW em uma determinada faixa de valores. Os arquivos criados, como o da listagem acima, podem ser visualizados sem o uso de programas específicos, o que poderia tornar mais flexível o seu uso.

Após a construção desta estrutura e visando avaliar sua eficiência para a obtenção de respostas necessárias à avaliação teórica da técnica, foi desenvolvido um novo *script* de computador para realizar buscas sobre esta base de dados segundo os critérios de faixa de pI e MW de peptídeos e proteínas, além da presença de cisteínas. Este tipo de busca é das mais simples que seriam utilizadas já que, visando uma avaliação, ainda seriam necessárias consultas envolvendo resumo de dados e estatísticas sobre todo o conjunto de peptídeos presentes. No entanto, os testes realizados com este último *script* mostraram certa lentidão, com tempos variando de 23 a 6 segundos para a obtenção de resultados. Também logo se verificou que, para cada novo tipo de consulta que se procurasse realizar visando uma forma diferente de extração de informações, seria necessário desenvolver um novo programa, o que acarretaria em consumo de muito tempo em desenvolvimento de *softwares*.

Visando solucionar estes problemas, buscou-se uma nova forma de estruturação dos dados, agora baseada num Sistema de Gerenciamento de Banco de Dados desenvolvido por terceiros. Assim, empregou-se o sistema MySQL, que permite a organização de dados em tabelas, criando automaticamente diferentes índices, e que permite o acesso às informações utilizando a linguagem de consulta SQL (*Structured Query Language*). O uso de SQL evita a necessidade de se criar novos *script*s para cada novo tipo de consulta; neste caso, o próprio sistema possui mecanismos para a busca dos dados.

Deste modo, um banco de dados foi criado com o MySQL, organizando as informações sobre peptídeos em duas tabelas distintas: "proteínas" e "peptídeos". Cada entrada (ou linha) presente nestas tabelas iriam guardar um determinado conjunto de dados (ou campos) conforme a especificação mostrada nas Tabelas 3.4 e 3.5.

Depois de definida a estrutura de dados no MySQL, o conjunto de ORFs e peptídeos da *Xylella fastidiosa* foi nela armazenado utilizando ferramentas internas do MySQL

Tabela 3.4: Estrutura dos tipos de campos das entradas da tabela "proteínas". Cada entrada desta tabela representa uma ORF do genoma da *Xylella fastidiosa*.

Nome do campo	Tipo de dado aceito	Descrição
Protein Accession Code	Texto	Identifica cada proteína entrada
MW	Numérico	Calculada usando a massa média e não a monoisotópica.
pI	Numérico	Calculado sem considerar as cisteínas.

Tabela 3.5: Estrutura dos tipos de campos das entradas da tabela "peptídeos". Cada entrada desta tabela representa um peptídeo e faz referência à proteína que a originou na tabela "proteínas".

Nome do campo	Tipo de dado aceito	Descrição
Id	Numérico	Identifica unicamente cada peptídeo
Id_Protein	Texto	Identifica a proteína que originou o Peptídeo. Referência a "Protein Accession Code" na tabela "proteínas"
MW	Numérico	Calculada usando a massa monoisotópica.
PI	Numérico	Calculado sem considerar as cisteínas.
Ini	Numérico	Posição de início do peptídeo na seqüência de sua proteína
Fim	Numérico	Posição final do peptídeo na seqüência de sua proteína.
Cisteínas	Numérico	Número de cisteínas presente no peptídeo
Seqüência	Texto	Seqüência de aminoácidos que forma o peptídeo.

e também o *script* para a digestão teórica. Consultas equivalentes às realizadas anteriormente foram concluídas em tempos inferiores a 1 segundo, demonstrando uma maior eficiência para a nova forma de organização. Uma grande variedade de consultas pôde ser aplicada sobre os registros de peptídeos utilizando-se a SQL, sendo que os principais resultados são apresentados na Seção 3.2.4 adiante.

3.2.3 Resultados de visualização

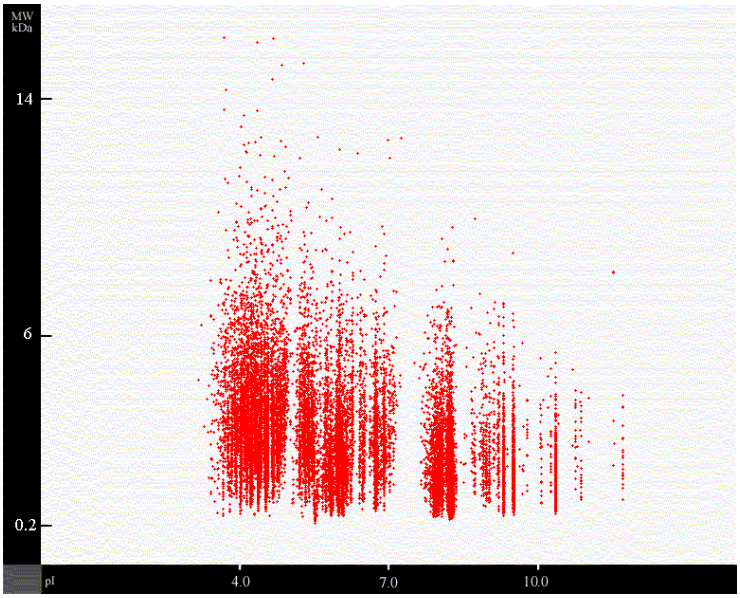


Figura 3.1: Distribuição de todos peptídeos (sem separação de MW de proteínas) numa 2DE teórica. Cada ponto vermelho representa um peptídeo. Utilizou-se uma escala de MW de 0 a 20kDa.

Resultados da visualização de peptídeos podem ser observados nas Figuras 3.1 e 3.2. A Figura 3.1, exibindo todo o conjunto de peptídeos (considerando-se inclusive aqueles que seriam obtidos com uma perda de quebra por tripsina) deixa evidente que há uma maior concentração de peptídeos com peso inferior a 6kDa, havendo mesmo alguns com pesos que chegam próximo a 19kDa. Peptídeos muito grandes podem ter aparecido por uma combinação de fatores como ocorrência da perda de quebra por lisina (aqui mantida no máximo de uma), ocorrências de prolinas nos lados C-terminais dos aminoácidos que indicam a quebra (lisinas e argininas).

Dos mais interessantes fatos que podem ser vistos na Figura 3.1, está a ocorrência de faixas, ou listras, verticais devido a existência de grandes quantidades de peptídeos com valores de pI próximos. Pode-se supor assim que o fator do pI não será tão determinante como critério de busca contra o banco de dados se comparado com o MW. Isto fica claro observando a Figura 3.2, onde são exibidos somente os peptídeos com peso entre 2.630 Da e 2.650 Da, numa escala compatível. Nela, mesmo exibindo-se uma

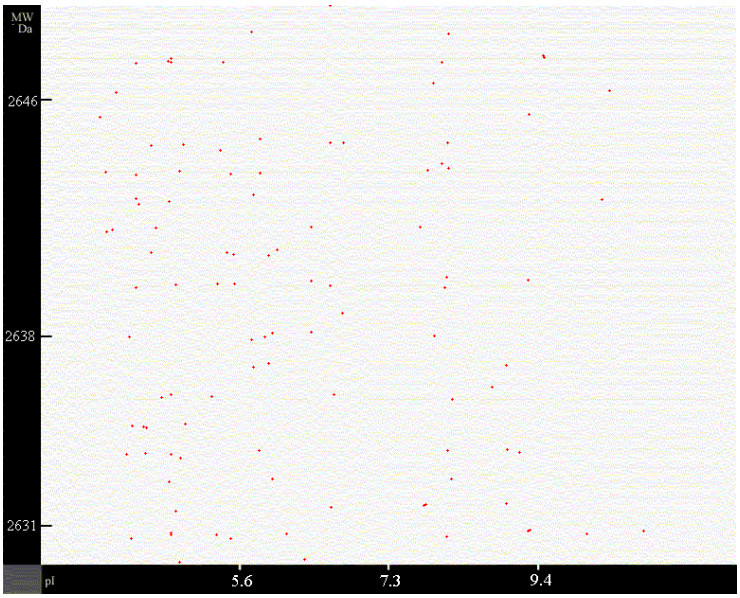


Figura 3.2: 2DE teórica de peptídeos, dentro de uma escala de MW de 2.630 a 2.650 Da.

extensão de pI que contém todos peptídeos presentes na Figura 3.1, conseguiu-se uma drástica redução do número deles. Assim, dada a importância do MW dos peptídeos, o fato das medidas dos pesos dos peptídeos poderem ser feitas com precisão de 70ppm será muito importante. De outro modo, o fato do pI não ser tão significativo pode ser utilizado para se trabalhar com faixas mais largas deste critério. O que é um fator positivo, tendo em vista as discrepâncias que poderão ser observadas entre estimativas teóricas e medidas experimentais de pI.

3.2.4 Resultados de consultas estatísticas sobre banco de dados

Nesta seção apresentamos resultados obtidos com as consultas em si ao banco de dados de peptídeos.

Foi indicado (Seção 1.1.4) que a separação dos peptídeos que contivessem cisteínas levaria a uma grande redução no número deles, contribuindo para o processo de identificação. Neste sentido, isto também pôde ser verificado em nossos testes, conforme demonstra a Tabela 3.6.

Pouco mais de 13% dos peptídeos gerados (incluindo a ocorrência de uma falha de quebra) apresentam alguma cisteína e assim podem contribuir para o processo de identificação. Neste caso, é relevante verificar se, com a exclusão de um número tão elevado de peptídeos, muitas proteínas não acabam sendo excluídas, o que seria algo negativo, pois elas estariam impossibilitadas de serem identificadas pela técnica. No entanto, conforme indica o gráfico da Figura 3.3, isso não ocorre.

Vê-se que mais de 83% das proteínas apresentam cisteínas. Tendo por base a

Tabela 3.6: Número de peptídeos agrupados segundo o número de cisteínas presentes em suas seqüências de aminoácidos, para as ORFs da *Xylella fastidiosa*. Destaca-se que, dos 149.946 peptídeos, 129.617 (86,4%) não apresentam cisteínas.

Cisteínas	Peptídeos	%
0	129.617	86,40%
1	16.010	10,70%
2	3.368	2,20%
3	697	0,50%
4	171	0,10%
5	52	0,03%
6	14	0,009%
7	7	0,005%
8	4	0,003%
9	3	0,002%
10	1	0,001%
14	2	0,001%

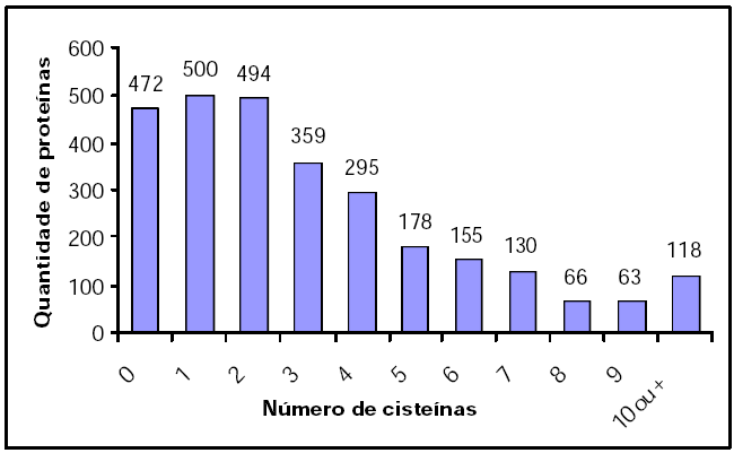


Figura 3.3: Gráfico da quantidade de proteínas agrupadas pelo número de cisteínas em suas seqüências. Do total de 2.830 proteínas, 472 delas (16,7%) não apresentam alguma cisteína na seqüência de aminoácidos.

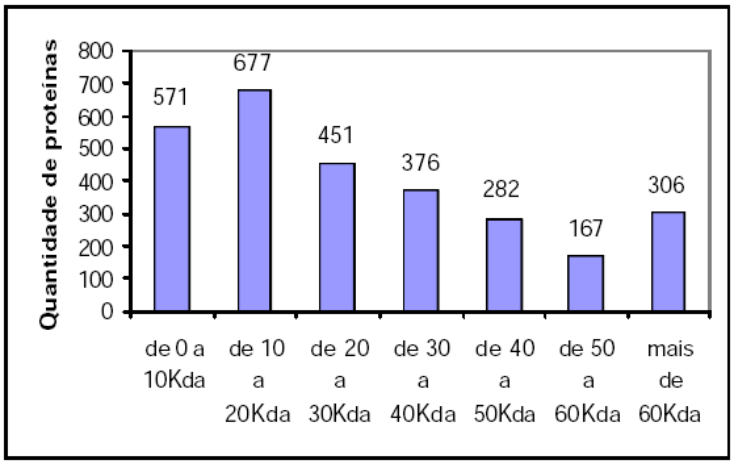


Figura 3.4: Distribuição dos valores de MW para as proteínas em faixas de 10kDa.

grande quantidade delas que são excluídas nas análises por 2DE (Gygi e Aebersold, 2000), pode-se concluir que esta técnica pode levar a um processo de identificação mais global do que o atualmente existente, desde que efetivamente realize a identificação unívoca de proteínas, algo analisado adiante. Procurando analisar a seletividade dos critérios de restrição de busca ao banco de dados, decidiu-se dividir o espaço de possibilidades para os valores de MW e pI dos peptídeos e MW de suas proteínas em faixas de valores; e então fazer uma contagem do número de peptídeos que há dentro de cada combinação (ou região) destas faixas. Assim, por exemplo, seria contado o número de peptídeos dentro da região onde seu MW está entre 1500,0 e 1500,5Da, o valor de pI está entre 4,5 e 5,0 e o MW de proteína se encontra entre 20kDa e 30kDa. Nestes testes somente foram utilizados os peptídeos com alguma cisteína, num total de 20.329, refletindo os passos experimentais de separação os peptídeos que foram marcados pelo reagente ICAT. As faixas de valores foram escolhidas segundo as especificações abaixo:

- A partir da distribuição das massas teóricas de proteínas, gráfico da Figura 3.4, foi adotado o estudo do critério de MW para proteínas em 3 faixas: de 0 a 20kDa (total de 1.248 proteínas), de 20 a 50kDa (1.109 proteínas) e maiores que 50kDa (473 proteínas).

- Já no estudo do critério de pI de peptídeos, a extensão de 0 a 14 unidades de pH foi dividida em faixas com tamanho de uma unidade (total de 14 faixas), de modo a serem até excessivamente largas para que possam superar erros nas medidas de pI dos peptídeos. Além disso, também foram testadas faixas com 0,5 unidades de largura (total de 28 faixas).

- Como um espectrômetro de massa pode fazer medidas com precisão de 70ppm, a extensão de 0 a 20kDa, foi dividida em faixas de valores com tamanho não superior a 100ppm em relação ao critério de massa dos peptídeos, para se garantir

uma maior faixa de segurança nos resultados. Assim, de acordo com o tamanho dos peptídeos, foram utilizadas faixas conforme o especificado na Tabela 3.7.

Tabela 3.7: Tamanho das faixas de MW utilizadas para o critério de MW de peptídeo.

Para a extensão de MW de peptídeos:	Tamanho da faixa utilizada
até 1kDa	0,1 Da
entre 1 e 5kDa	0,5 Da
entre 5 e 10kDa	1,0 Da
após 10kDa	2,0 Da

Comparações entre valores experimentais e teóricos de pI e MW de proteínas por técnicas de eletroforese, em geral, levam a significativas diferenças, algo que já foi verificado na Seção 2.1.3. Por isso, buscou-se deixar as faixas de busca suficientemente largas. Levando em conta as condições mais favoráveis para estimativas experimentais de MW e pI na forma como poderão ser feitas nos experimentos pela abordagem estudada aqui, o uso destas faixas mais largas dão maior crédito aos estudos teóricos realizados. Essas "condições mais favoráveis" podem ser explicitadas como:

- Verificação experimental de pI de peptídeos e não proteínas. Nos peptídeos pequenos há uma impossibilidade de formação de estruturas terciárias e as cargas dos aminoácidos estarão mais expostas ao meio (fator assumido no cálculo teórico).
- Medida de MW de peptídeos por espectrômetro de massa. O que, devido a alta precisão, facilita a correlação com estimativas teóricas.
- O uso de SDS-Page na separação de faixas de proteínas é mais global do que o de 2DE, permitindo estudo de maior número de proteínas.

Os resultados das contagens do número de peptídeos em cada região estão nos gráficos das Figuras 3.5, 3.6, 3.7 e 3.8, adiante. Nelas, foi totalizado o número de regiões com um mesmo número de peptídeos. Pode-se observar que as colunas dos gráficos que indicam o número de regiões onde somente são encontrados um único peptídeo são sempre maiores que as demais, numa proporção sempre entre 75% e 93% das regiões. Isto indica que, segundo previsões teóricas, a técnica pode ser utilizada para se encontrar a proteína originária, e assim identificá-la, de um grande número de peptídeos. No caso das Figuras 3.5, 3.6 e 3.7, um total de 15.430 peptídeos (11.129 na Figura 3.8) se encontram sozinhos numa região considerada e, assim, seriam identificáveis segundo a separação de regiões realizada nestes testes, onde buscou-se realizar uma avaliação global da técnica abrangendo todos os peptídeos. Entretanto, a técnica tem o potencial de tornar esse número maior tendo em vista que, não mais numa análise global, mas na caracterização de peptídeos individualmente, os critérios, sobretudo o de faixa de MW de peptídeo, seriam aplicados em torno dele, resultando numa seleção mais fina.

Cabe também ressaltar que a impossibilidade de identificar a proteína de um determinado peptídeo, por este ocorrer numa região com um ou mais outros peptídeos,

não inviabiliza a identificação da própria proteína, tendo em vista que ela pode ter originado outros peptídeos com cisteína e que estes sejam identificáveis. Isto, minimamente, pode ser verificado pelo seguintes números: os 11.129 peptídeos, que ocorreram unicamente no teste representado na Figura 3.8, estão relacionados a 2.193 proteínas (ORFs), numa relação média de 5,1 peptídeo para cada proteína.

Este número de 2.193 proteínas identificáveis no teste da Figura 3.8 é de grande interesse também, tendo em vista que representa mais de 77% das 2.830 ORFs do genoma da *Xylella fastidiosa*. O que evidencia, teoricamente, que esta técnica é mais global do que as técnicas de identificação que utilizam eletroforese 2D.

É importante comentar que, no gráfico da Figura 3.8, vê-se o resultado dos testes sem a separação de uma faixa de MW de proteínas, o que equivale experimentalmente a não se realizar a SDS-Page (Figura 1.4). Neste caso, o processo torna-se menos seletivo, visto que a proporção de regiões com um único peptídeo, 75% ou 82%, é menor do que as obtidas com a separação de faixas, Figuras 3.5, 3.6 e 3.7. Por um lado, isto dá mais força a afirmação anterior sobre a globalidade da técnica, visto que há possibilidade de obter-se ainda melhores resultados utilizando separação por faixas de MW de proteínas. E por outro lado, vê-se, ainda pelo número de 2.193 ORFs identificáveis na Figura 3.8, que a eliminação do passo com SDS-Page pode levar a bons resultados, mesmo sendo menos seletiva. Devido a maior facilidade experimental que a eliminação da SDS-Page resulta, uma boa opção talvez seja o uso combinado de experimentos com e sem ela.

Como era de se esperar, os testes realizados sobre faixas de pI de 0,5 unidades originaram resultados melhores do que os sobre faixas de 1,0 unidade. Nesta comparação, com as faixas de 0,5 un. observa-se que sempre se atinge uma melhor seletividade, com um aumento da proporção de regiões com um único peptídeo, e uma conseqüente diminuição daquelas onde ocorrem 2 ou mais peptídeos, efeito desejado. Entretanto, esta melhora não é muito acentuada. Nesse sentido, verifica-se que mesmo o número total de regiões (dado que consta nas legendas das Figuras) não aumenta mais que 10% com o uso de faixas de 0,5 un., sendo que, como é uma faixa com metade do tamanho da anterior, não seria estranho um aumento da ordem de 100% no número de regiões. Este resultado é refletido também no aspecto das 2D teóricas das Figuras 3.1 e 3.2, onde observa-se a ocorrência de listras verticais, resultantes de peptídeos com pI muito próximos.

Cabe ainda ressaltar a ocorrência de algumas regiões com um número elevado de petídeos, até 56 na Figura 3.8. Observou-se a presença de vários peptídeos pequenos com seqüência de aminoácidos muito parecida. Exemplos destes encontram-se na Tabela 3.8, em todos os casos os peptídeos são originários de diferentes proteínas.

3.2.5 Ferramenta de Busca

Algoritmo para busca de peptídeos

Anteriormente ao processo de implementação da ferramenta de busca de peptídeos à base de dados, procurou-se estudar formas de fazê-lo visando uma forma eficiente de acesso aos dados e também o levantamento de uma pontuação (*score*) entre possíveis peptídeos localizados.

Tabela 3.8: Exemplos de peptídeos com múltiplas ocorrências na digestão teórica realizada.

Seqüência de aa.	pI teórico	MW teórico	Número de ocorrências
"CR" ou "RC"	9,75	719,35	56
"CK"	8,75	691,34	24
"R" após as diferentes permutações de "I" (ou "L") com "C"	9,75	832,43	16

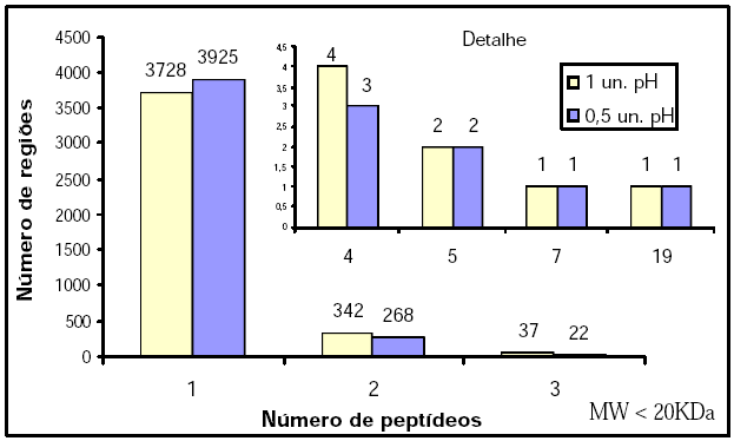


Figura 3.5: Gráfico do total de regiões (onde ocorreram ao menos um peptídeo) pelo número de peptídeos dentro delas. Aqui, foi considerada a faixa de MW de proteínas menores que 20kDa, ou seja, foram utilizados peptídeos originários de proteínas com massa inferior a 20kDa. Também são apresentados os resultados com o uso de faixas de pI com largura de 1 unidade e 0,5 unidades. O número de regiões onde há 4 ou mais peptídeos são apresentados no detalhe. Cabe ressaltar que, das 4.115 regiões resultantes, 3.728 (91%) apresentaram um único peptídeo, no caso em que foram testadas faixas de pI de largura 1un. No caso de faixas com largura de 0,5 un., 3925 (93%) das 4.222 regiões apresentaram somente um peptídeo.

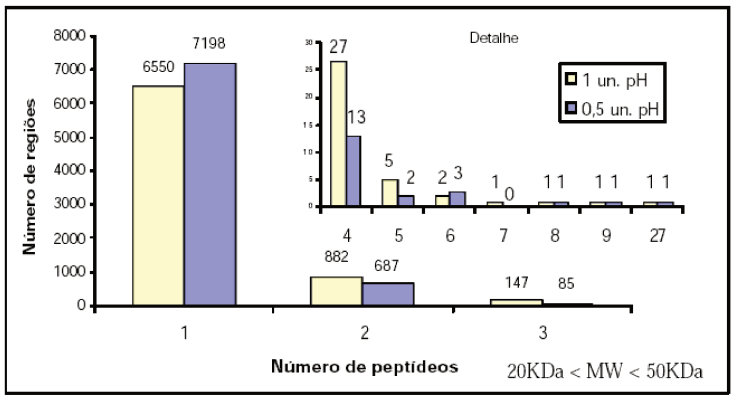


Figura 3.6: Gráfico do total de regiões (onde ocorreram ao menos um peptídeo) pelo número de peptídeos dentro delas. Aqui, foi considerada a faixa de MW de proteínas entre 20kDa e 50kDa, ou seja, foram utilizados peptídeos originários de proteínas com massa entre 20kDa e 50kDa. Também são apresentados os resultados com o uso de faixas de pI com largura de 1 unidade e 0,5 unidades. No caso com uso de faixas de pI em 1 un., das 7.617 regiões resultantes com ao menos um peptídeo, 6.550 (86%) apresentam só um peptídeo. No caso de faixas de 0.5un., 7.198 (90%) das 7.991 regiões apresentam somente um peptídeo também.

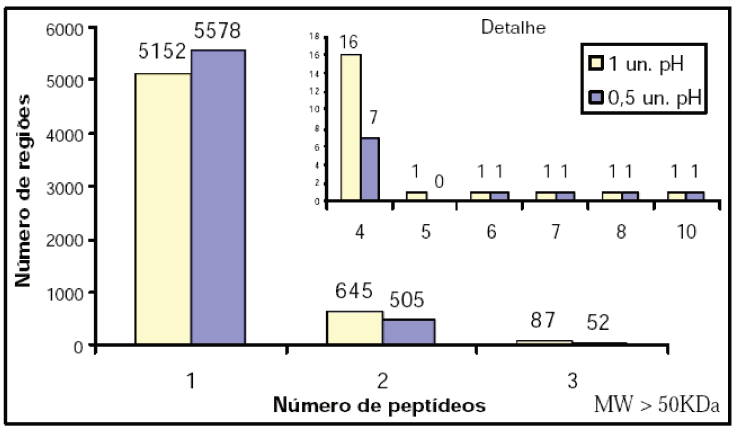


Figura 3.7: Gráfico do total de regiões (onde ocorreram ao menos um peptídeo) pelo número de peptídeos dentro delas. Aqui, foi considerada a faixa de MW de proteínas maiores que 50kDa, ou seja, foram utilizados peptídeos originários de proteínas com massa maior que 50kDa. Também são apresentados os resultados com o uso de faixas de pI com largura de 1 unidade e 0,5 unidades. No caso com uso de faixas de pI em 1 un., 5.152 (87%) das 5.905 regiões resultantes apresentam um peptídeo apenas. Com faixas de pI de 0,5 un, são 5.578 (91%) das 6.146 regiões exibem um único peptídeo.

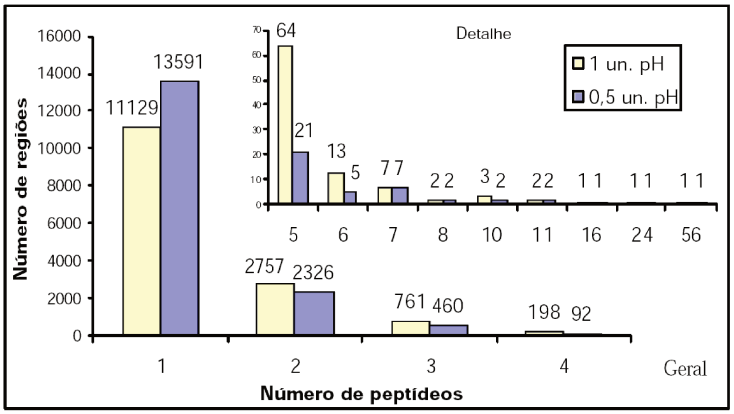


Figura 3.8: Gráfico do total de regiões (onde ocorreram ao menos um peptídeo) pelo número de peptídeos dentro delas. Aqui, não foi considerada a faixa única de MW para as proteínas, assim foram utilizados todos os peptídeos, não importando a massa de sua proteína originária. Também são apresentados os resultados com o uso de faixas de pI com largura de 1 unidade e 0,5 unidades. Destaca-se que, das 14.939 regiões resultantes, 11.129 (75%) apresentam um único peptídeo, no caso em que se usa faixas de pI de 1un. E 13.591 (82%) das 16.511 regiões geradas no caso de faixa de 0,5 un. para pI apresentam somente um peptídeo.

Devido ao uso de massas de peptídeos e proteínas como critérios de busca, uma referência natural foram os algoritmos existentes para identificação de proteínas através de *peptide mass fingerprinting*. Pois eles também procuram correlacionar massas de peptídeos medidas experimentalmente com massas calculadas de peptideos derivados de proteínas armazenas numa base de dados (Henzel *et al.*, 1993; James *et al.*, 1993; Pappin *et al.*, 1993). O processo de identificação por *peptide mass fingerprinting* pode ser dividido em duas fases, segundo a descrição abaixo (Patterson *et al.*, 2001):

- Métodos experimentais:
  1. Partir da proteína intacta: por exemplo, separada em gel ou cromatograficamente purificada.
  2. Obtenção de peptídeos por digestão enzimática.
  3. Medida de massas de peptídeos com espectrometria de massa (MS).
- Métodos computacionais:
  1. Partir de uma base de dados de seqüência de proteínas, junto com tradução de base de dados de seqüência de nucleotídeos.
  2. Busca de massa de peptídeos
     (a) Para cada entrada, calcular as massas dos peptídeos obtidos segundo uma determinada especificidade de enzima.

(b) Correlacionar as massas observadas de peptídeos com massas medidas experimentalmente.

(c) Ordenar resultados segundo melhor correlação ou somente aqueles que atingem uma significância predeterminada, baseada no número de peptídeos correlacionados.

Desta descrição, vê-se que a fase experimental se assemelha com o descrito anteriormente para o método de identificação deste projeto (Figura 1.4), de onde o resultado final é uma lista de massas de peptídeos determinadas por MS. Dentro da fase computacional, destaca-se que o Item 2a, de obtenção de massas de peptídeos, se assemelha com a fase anterior deste projeto, quando inclusive houve a inserção de informações sobre peptídeos teóricos numa base de dados MySQL (Seção 3.2.2). Restando assim os itens de correlação de massas, levantando *scores*.

Devido a implementação do banco de peptídeos em MySQL, os recursos de consulta e indexação deste sistema de gerenciamento de dados podem ser utilizados para a obtenção de entradas segundo critérios estabelecidos sobre os dados nele armazenados. Assim, podem ser utilizadas expressões como a da listagem abaixo para a obtenção da lista de peptídeos que se correlacionam à massas medidas experimentalmente por MS:

```
SELECT proteins.*, peptides.*
FROM proteins INNER JOIN peptides
WHERE
proteins.mw >= 10000 and proteins.mw <= 20000 and
peptides.pi >= 3 and peptides.pi <= 8 and
peptides.mw >= 1230 and peptides.mw <= 1231
```

Neste exemplo, são selecionados peptídeos de pI teórico entre 3 e 8, MW entre 1230 e 1231 (equivalente a uma massa experimental de 1230,5 $\pm$ 0.5Da) e MW da proteína original (intacta) entre 10kDa e 20kDa.

Para obtenção de uma pontuação entre proteínas localizadas nas buscas de *peptide mass fingerprinting*, é possível calcular a porcentagem do número massas experimentais de peptídeos em comum com as dos peptídeos da proteína localizada na base de dados[5] (Henzel *et al.*, 1993; Wilkins e Williams, 1997). Deste modo, uma melhor pontuação estaria relacionada a um maior número de massas de peptídeos medidas que podem ser localizadas entre os fragmentos de uma proteína da base de dados. Logo, pode-se ver que tal procedimento não seria aplicável à estratégia em estudo neste projeto. Já que, aqui, trabalha-se com a massa medida de um único peptídeo, o que faria com que qualquer proteína localizada obtivesse a mesma pontuação máxima, caso se empregasse tal forma de pontuação.

Este tipo de dificuldade é intrínseca a qualquer estratégia desenvolvida para *peptide mass fingerprinting*, uma vez que esta busca a obtenção de um maior número de medidas (maior número de peptídeos) que, assim, possbilitam a identificação da proteína.

No entanto, pôde-se encontrar uma alternativa através do emprego do algoritmo Mowse[6] [7] (Pappin *et al.*, 1993). Pois ele se baseia na freqüência de distribuição dos

[5] Conforme http://ca.expasy.org/tools/peptident-doc.HTML, Ago/2005.

[6] http://prospector.ucsf.edu/ucsfHTML4.0/instruct/fitman.htm, Ago/2005.

[7] http://www.hgmp.mrc.ac.uk/Bioinformatics/Webapp/mowse/mowsedoc.HTML, Ago/2005.

peptídeos da fonte de dados (Pappin *et al.*, 1993), de modo que, mesmo no caso da ocorrência de proteínas com um único peptídeo correlacionado, pode-se dar pesos diferentes para tais peptídeos, obtendo pontuações diferentes.

Para computar a pontuação no Mowse, as proteínas (intactas) da base de dados são agrupadas em intervalos de MW de 10kDa. Dentro de cada intervalo, os peptídeos, originados da digestão teórica das proteínas presentes, são agrupados em células de 100Da segundo o MW destes peptídeos e, para cada célula, é contado o número de peptídeos presentes. Então, é calculada a freqüência de cada célula como a razão entre a quantidade de peptídeos nela e o total da soma do número de peptídeos em cada intervalo de 10kDa de proteína. Finalmente, as freqüências são normalizadas dividindo o valor antigo de freqüência pela maior freqüência encontrada, para cada intervalo de 10kDa, o que gera freqüências normalizadas entre 0 e 1.

No Mowse, durante a busca de uma proteína, cada massa medida de peptídeo que correlacionar com um peptídeo teórico da proteína analisada contribui, para a pontuação, com o valor de freqüência normalizada correspondente. No caso de mais de um peptídeo correlacionado, as freqüências são multiplicadas. O produto final de freqüências (P_N), assim obtido, é invertido e normalizado para um MW médio de proteínas de 50kDa, obtendo a pontuação final (S) conforme a Equação 3.1:

$$S = \frac{50000}{P_N \times H} \tag{3.1}$$

Onde H é o peso molecular teórico da proteína analisada, a qual será atribuída a pontuação. As proteínas com menores pontuações são selecionadas e ordenadas.

Aplicando este método na estratégia estudada neste projeto, o produto P_N sempre será resultado de um único peptídeo correlacionado, o que diminui a importância desta pontuação se comparado ao que é observado em *peptide mass fingerprinting*. No entanto, tal forma de pontuação foi utilizada na implementação final do mecanismo de busca com o objetivo de fornecer uma informação extra aos pesquisadores que estejam utilizando-o, facilitando a interpretação de resultados.

Outra informação calculada que pode ser utilizada para avaliação dos resultados é a diferença, em Daltons, entre a massa medida de peptídeo fornecida como parâmetro de busca e a massa do peptídeo localizada no banco de dados.

Cabe ressaltar que o método de identificação de proteínas aqui estudado visa uma busca de peptídeos que obtenha uma única possível proteína como resultado. Esta é a situação em que melhor se observa a eficiência do método e que se buscou estudar na fase anterior deste projeto. Com a obtenção de uma única entrada da base de dados como resultado, as informações de pontuação ou diferença do valor teórico e experimental não seriam utilizadas como discriminante entre melhor e pior resultado, mas como parâmetro para se atribuir maior ou menor significância ao resultado final da busca.

Implementação de Ferramenta de Busca

A ferramenta de busca, a qual será referida por FindPep, pode ser dividida em 3 componentes principais:

1. Interface de entrada de parâmetros.
2. Programa de busca em banco de dados com base nos parâmetros fornecidos.
3. Interface de Exibição de resultados da busca.

Visando desenvolver um software que pode ser utilizado via Internet, estes componentes foram criados utilizando tecnologias específicas para este fim, como páginas HTML e *scripts* armazenados em servidor *web*.

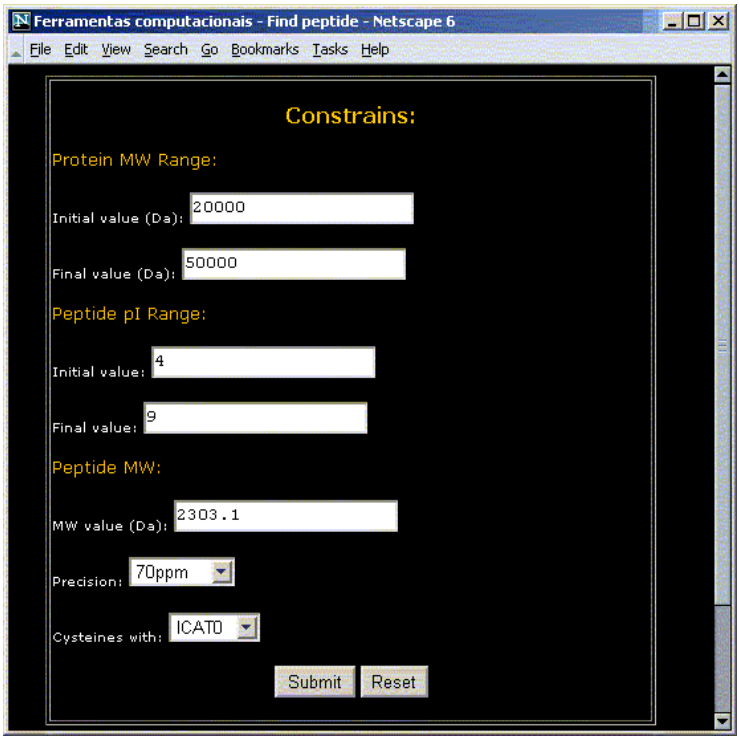


Figura 3.9: Formulário para entrada de parâmetros de busca. Acessível em http://www.proteome.ibi.unicamp.br/tools/findpep (Out/2006).

A interface da ferramenta foi implementada através de um formulário HTML, Figura 3.9, com os campos necessários para cada parâmetro em uso no algoritmo de busca:

Faixa de MW para proteína com um campo para o MW de início da faixa e outro para o MW final. Faixa de pI para o peptídeo, também com um campo para valor inicial e outro para valor final, determinando a faixa de valores.

Valor de MW para o peptídeo dado obtido de espectrômetro de massa.

Precisã utilizada para a seleção de peptídeos com valores entre 50ppm e 150ppm. Na busca, uma massa experimental de peptídeos, F_{MW}, é correlacionada à uma massa P_{MW} de peptídeo da base de dados se o seguinte critério é satisfeito: $F_{MW} \times (1-Precisão) \leq P_{MW} \leq F_{MW} \times (1+Precisão)$.

Reagente ligado às cisteínas com as opções de ICAT0 (ICAT sem deutérios), ICAT8 (ICAT com 8 deutérios) e PEO (também pode ser usada para seleção de peptídeos com cisteína). Conforme o tipo de reagente ligado às cisteínas, têm-se um MW teórico diferente para os peptídeos da base de dados. Assim, este parâmetro presa ser fornecido.

Os valores inseridos no formulário são enviados, com a submissão, ao servidor, onde são tratados pelo *script* findpep.php. Este *script* corresponde ao segundo componente listado anteriormente. Através de uma conexão com o servidor de MySQL presente na máquina, ele pode ter acesso à base de dados com tabelas de informações de proteínas e peptídeos da bactéria *Xylella fastidiosa*, de onde seleciona as entradas que correspondem aos parâmetros fornecidos pelo formulário de parâmetros. Tal seleção envolve a criação de uma expressão de consulta como a da listagem 1 e sua execução pelo MySQL, o que retorna ao *script* os dados correspondentes: lista de peptídeos com pI e MW na faixa tolerada e correspondentes proteínas que originam tais peptídeos. Cabe ressaltar que são selecionados os peptídeos que contém cisteínas.

Como parte do processo de levantamento de uma pontuação entre as entradas retornadas, mantém-se armazenada na base de dados uma tabela com as freqüências normalizadas dos diferentes intervalos de MW. Tal tabela é consultada após a obtenção da lista de peptídeos correlacionados aos dados experimentais. Também é calculada a diferença entre as massas dos peptídeos localizados e da massa de peptídeo fornecida no formulário de parâmetros.

Os resultados são retornados ao usuário através de uma página HTML onde podem ser conferidos os parâmetros utilizados para a restrição à busca e também uma tabela com os peptídeos localizados, Figura 3.10.

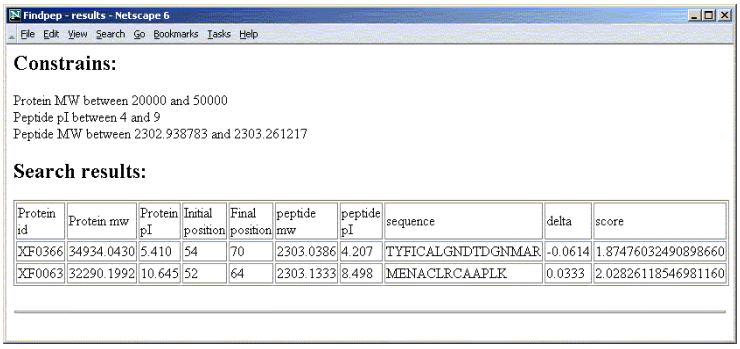


Figura 3.10: Página de resultado de busca de peptídeos

A tabela de peptídeos localizados apresenta as seguintes colunas:

Protein id número de acesso à proteína intacta que originou o peptídeo localizado.

Protein MW peso molecular teórico da proteína, calculado usando pesos moleculares médios de aminoácidos.

Protein pI ponto isoelétrico teórico da proteína, calculado considerando que as cisteínas não se ionizam.

Initial position e Final position número, dentro da seqüência da proteína intacta, do primeiro e do último aminoácido do peptídeo localizado.

Peptide MW peso molecular teórico do peptídeo localizado, calculado usando pesos isotópicos de aminoácidos.

Peptide pI ponto isoelétrico teórico do peptídeo.

Sequence seqüência de aminoácidos do peptídeo.

Delta diferença entre o MW teórico do peptídeo e o MW experimental fornecido como parâmetro.

Score pontuação do peptídeo, seguindo o esquema de pontuação do algoritmo Mowse.

Testes

Até o momento não é possível realizar testes experimentais plenos da ferramenta FindPep uma vez que algumas das técnicas para obtenção de todos os parâmetros de restrição às buscas e, sobretudo quanto ao uso combinado delas, ainda está sob padronização e verificação. No entanto, tais testes não haviam sido previstos no plano inicial e, mesmo anteriormente a eles, deveriam ser realizados testes que visassem observar o correto funcionamento da ferramenta, sobre condições bem conhecidas e sobre dados armazenados na base de dados.

Assim sendo, foram selecionadas as proteínas da Tabela 3.9 para a realização de testes. A digestão teórica delas origina 33 peptídeos, que constam na Tabela 3.10. Cada um destes foi submetido pelo formulário de entrada da FindPep, utilizando os seguintes critérios de restrição:

- Para MW de proteína, utilizou-se uma das faixas 0-20kDa, 20-50kDa e maior que 50kDa, conforme qual delas pertencia a proteína.
- A faixa de pI de peptídeo foi tal que o pI inicial da faixa era uma unidade inferior ao pI teórico de cada peptídeo e o pI inicial, uma unidade superior ao pI teórico.
- Para o MW de peptídeo, foi fornecido exatamente o valor de MW teórico de cada peptídeo testado, considerando um desvio de até 70ppm.

Como era de se esperar, em todos os testes, o peptídeo fornecido como parâmetro foi corretamente localizado. A última coluna da Tabela 3.10 expressa o número de peptídeos que foram obtidos, como resultado da busca, quando o peptídeo de cada linha foi submetido ao FindPep. Em 23 dos 33 testes, correspondendo a 70%, a ferramenta forneceu um único peptídeo como resultado da busca. Nos outros 10 testes, obteve-se 2 peptídeos, de proteínas diferentes daquelas que originaram o peptídeo em teste.

Tabela 3.9: Proteínas utilizadas em testes.

Número de identificação	MW	pI	Número de cisteínas presente
XF0290	100621,8	6,24	5
XF2394	30428,78	4,51	2
XF1502	11677,96	4,61	1
XF1548	54772,6	7,97	6

A partir deste momento, a ferramenta FindPep está disponível através do endereço eletrônico http://www.proteome.ibi.unicamp.br/tools/findpep/. Na medida que testes experimentais fiquem disponíveis eles serão utilizados em conjunto com o FindPep para a verificação do método de identificação de proteínas. Um passo importante para isso será confrontar os resultados deste método com o de outros já amplamente utilizados como, por exemplo, MS/MS, verificando se todos métodos levam a identificação da mesma proteína.

Tabela 3.10: 33 peptídeos originados da digestão teórica das proteínas da Tabela 3.9 e que contenham ao menos uma cisteína. Foi considerada a possibilidade da ocorrência de uma falha de quebra da enzima de digestão, lisina. A coluna "Cys" mostra o número de cisteínas na seqüência de aminoácidos do peptídeo. E a coluna "N" mostra o número de entradas retornada com uma busca no FindPep usando os dados destes peptídeos como parâmetros de restrição.

Proteína intacta	MW do peptídeo	pI do peptídeo	Seqüência de aminoácidos do peptídeo	Cys	N
XF0290	2474.2734	4,208	VVLQDFTGVPCVVDLAAMR	1	1
XF0290	3000.5598	4,429	VVLQDFTGVPCVVDLAAMRDAAIR	1	2
XF0290	3261.6560	5,836	DGAVVIAAITSCTNTSNPAVMFGAGLLAR	1	2
XF0290	3458.6560	4,137	ATIGNMAPEYGATCGIFPIDTESLNYLR	1	1
XF0290	3744.9365	8,748	DGAVVIAAITSCTNTSNPAVMFGA-GLLARNAVAK	1	1
XF0290	3871.8948	4,679	ATIGNMAPEYGATCGIFPIDTESLNYLRLSGR	1	1
XF0290	3902.9441	4,105	AEPDTEIAFMPARVVLQDFTGVPCVVDLAAMR	1	1
XF0290	4780.3877	4,137	AGLLNDLETLGFYVVGYGCTTCIG-NSGPLPPEVSAGIAK	2	2
XF0290	4884.5186	4,429	GQVDLDINGQTLQLKDGAVVIAAI-TSCTNTSNPAVMFGAGLLAR	1	1
XF0290	5359.6030	4,291	FVEFYGDGLAHLPLADRATIGNMA-PEYGATCGIFPIDTESLNYLR	1	1
XF0290	5727.8838	4,178	VVTDYLEKAGLLNDLETLGFYVVG-YGCTTCIGNSGPLPPEVSAGIAK	2	1
XF0290	6004.0498	4,317	AGLLNDLETLGFYVVGYGCTTCIG-NSGPLPPEVSAGIAKGDLVAAAVLSGNR	2	1
XF1548	1144.5984	8,720	VGCIPSK	1	1
XF1548	1387.7203	5,806	VACVDAALGK	1	2
XF1548	1628.8378	8,748	DGKPALGGTCLR	1	1
XF1548	1886.9958	8,718	VGCIPSKALLDSSR	1	1
XF1548	2069.1377	8,635	AAQLGLKVACVDAALGK	1	1
XF1548	2643.2971	6,261	ICHAHPTLSEAIHDAAMAVSK	1	1
XF1548	2755.4258	9,994	DGKPALGGTCLRVGCIPSK	2	1
XF1548	2799.3982	7,025	ICHAHPTLSEAIHDAAMAVSKR	1	1
XF1548	2998.5476	8,560	VACVDAALGKDGKPALGGTCLR	2	1
XF1548	3448.6399	4,636	GQIVVDEHCHTGVDGVWAIGDCVR	2	1
XF1548	4183.0298	5,741	GQIVVDEHCHTGVDGVWAIGDCVRGPMLAHK	2	1
XF1548	4872.3818	4,508	GLLADGTGVQLNERGQIVVDEHCHTG-VDGVWAIGDCVR	2	1
XF1548	6072.0732	5,235	VLGLHLIGVNVSELVHEGVLAMEFSG-SADDLARICHAHPTLSEAIHDAAMAVSK	1	1
XF1502	1944.9648	4,137	ITVEDCLEVVNNR	1	1
XF1502	2303.1436	4,677	MARITVEDCLEVVNNR	1	1
XF1502	2981.4734	4,407	ITVEDCLEVVNNRFELVMMASK	1	2
XF2394	1409.6506	8,500	MPQCGFSAK	1	2
XF2394	1878.9233	7,016	ETPLAFLCHHGGR	1	2
XF2394	2729.4146	6,967	AALEQLPKETPLAFLCHHGGR	1	2
XF2394	2918.4431	7,119	ETPLAFLCHHGGRSLQAAEHFR	1	2
XF2394	3984.9900	4,749	MPQCGFSAKAAGILQALGVEYAHVN-VLDDQEIR	1	2

Capítulo 4

Ferramenta de análise por vias metabólicas

A última ferramenta desenvolvida iniciou com a formação de uma base de dados sobre vias metabólicas e então o desenvolvimento de ferramentas para consulta a esta base.

4.1 Obtenção de uma representação para os dados

Podemos definir "via" metabólica como a seqüência de reações químicas que ligam um composto inicial à um composto terminal, sendo que eventuais desvios para outro composto terminal, a partir do mesmo inicial, podem ser tratados como nova via (Michal, 1998). Sendo assim, dois elementos importantes para serem representados nas estruturas de dados para armazenamento de vias metabólicas reações químicas e compostos necessários a elas: substratos e enzimas. Através da KEGG (Seção 1.3.1), temos acesso a descrições de vias metabólicas com figuras com sua representação gráfica (arquivos em formato GIF). Assim, estas devem ser representadas também tendo em vista poder-se agrupar reações e enzimas que participam em vias comuns.

Por motivos de simplificação na computação de caminhos de vias metabólicas, torna-se mais eficiente representar as reações como relações binárias de substrato-produto (Kanehisa, 2000), onde dois compostos (substratos) são associados entre si através de uma enzima. Neste caso, reações que envolvem múltiplos produtos ou substratos são decompostas, por fins de representação, em todos pares substrato/produto possíveis (Kanehisa, 2000). Os bancos de dados obtidos pelo KEGG são em formato texto puro com campos nomeados em cada linha como a listagem abaixo com as duas primeiras entradas do banco de dados de reações:

```
ENTRY       R00001
NAME        Polyphosphate polyphosphohydrolase
DEFINITION  Polyphosphate + H2O <=> Oligophosphate
EQUATION    C00890 + C00001 <=> C02174
ENZYME      3.6.1.10
```

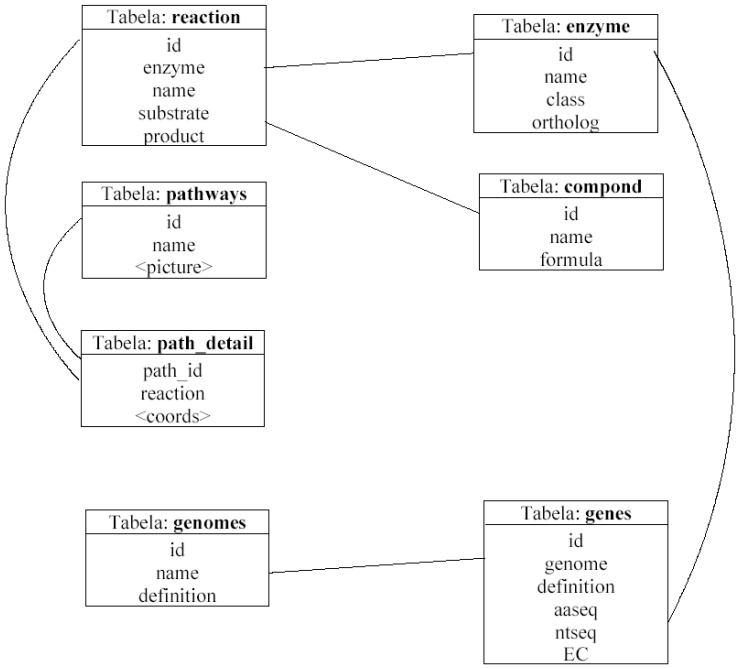


Figura 4.1: Representação da estruturas das tabelas para armazenamento de dados sobre vias metabólicas. Cada quadro indica uma tabela e possui uma lista dos campos da tabela. As linhas entre tabelas indicam relacionamentos (referências) entre tabelas, de modo que uma entrada pode recuperar outra entrada numa tabela referida.

```
///
ENTRY       R00002
NAME        Reduced ferredoxin:dinitrogen oxidoreductase (ATP-hydrolysing)
DEFINITION  16 ATP + 16 H2O <=> 8 e- + 8 H+ + 16 Orthophosphate + 16 ADP
EQUATION    16 C00002 + 16 C00001 <=> 8 C05359 + 8 C00080 + 16 C00009 +
            16 C00008
ENZYME      1.18.6.1
///
```

Assim, através de um programa de computador, eles precisariam ser acessados seqüêncialmente (do início ao fim a procura de entradas desejadas), tornando muito pouco eficientes sua manipulação. Além disso há replicação de dados entre bancos. Por exemplo, na listagem anterior os campos "Equation"e "Definition"contém a fórmula da reação em questão e códigos de identificação para os compostos envolvidos, ao mesmo tempo que o banco de dados de enzimas apresenta, para a mesma enzima 3.6.1.10, da primeira reação da listagem, uma descrição da mesma reação química e dos compostos. Esse tipo de redundância, útil num arquivo texto onde a busca de informações associadas a uma entrada já localizada é muito custosa, não seria aconselhável numa base de dados relacional, como a do MySQL a qual nos propomos a usar neste

projeto, devido ao risco de inconsistência entre as informações armazenadas. As tabelas do MySQL também permitem o uso de índices que tornam eficientes a busca por entradas nas tabelas, reduzindo tempo de processamento.

A estrutura básica para as tabelas que armazenarão os dados de vias, obtidas na KEGG, está representada na figura 4.1. Cada tabela pode armazenar uma seqüência virtualmente infinita de entradas com os campos que estão na descrição da tabela considerada. Entradas numa tabela podem fazer referência a dados em outra tabela com uso de campos com mesmos valores, representado na figura 4.1 por uma linha contínua.

Uma vez tendo observado isso, foram criadas rotinas de suporte em linguagem Java para converter os arquivos KEGG em MySQL. As rotinas desenvolvidas podem ser classificadas em dois tipos:

- Rotinas que percorrem um arquivo texto da KEGG, identificando os campos presentes em cada entrada do arquivo e computando dados estatísticos das tabelas.
- Rotinas para a conversão em si da base de dados. Estes programas, usam as informações obtidas anteriormente sobre os arquivos da KEGG para então realizar a identificação das diferentes entradas presentes e incluí-las diretamente na base de dados MySQL. Esta última etapa, envolvendo o uso da API (*Application Programming Interface*) de Java para conexão com servidor de dados.

Convém ressaltar que esses programas poderão ser utilizados em futuras atualizações do banco de dados a partir de novas versões da KEGG que ficarem disponíveis, automatizando o processo de conversão de dados.

Como resultado deste processo de execução dos programas de conversão, foi criada a banco de dados local com 10 tabelas de dados, conforme os dados apresentados na Tabela 4.1.

4.2 Ferramenta de consulta: ECPath

Uma vez definidas estruturas de dados e tendo os dados em si sobre vias metabólicas e reações químicas sido coletados, o próxima etapa, visando possibilitar o acesso a estes dados foi o desenvolvimento do programa *ECPath*.

A ferramenta *ECPath* tem por objetivo a automatização de alguns tipos de consultas comuns à base de dados de vias metabólicas. Por exemplo, através dos géis 2D utilizados em proteômica e da identificação dos *spots* presentes neles, chega-se a um conjunto de proteínas que foram expressas numa mesma situação ou momento celular. Estas proteínas, assim, podem desempenhar diversas relações entre si e, em especial, no caso destas proteínas serem identificadas como enzimas, podem estar relacionadas através da participação em mesmas vias metabólicas. A ECPath poderá explicitar esta relação buscando as vias metabólicas que apresentam as mesmas enzimas encontradas no experimento.

A ECPath também desempenha uma função auxiliar de localização das reações químicas que compõem vias metabólicas. Neste caso, no entanto, uma vez que o conjunto de reações presentes nas vias variam para organismos diferentes, optou-se por

Tabela 4.1: Lista de tabelas de dados MySQL criadas para armazenamento de dados sobre vias metabólicas. São apresentados detalhes sobre o número de registros e espaço físico ocupado na primeira versão realizada da conversão de dados da KEGG

Tabela Criada	Descrição	Número de registros presente	Espaço em disco ocupado
compound	Lista de compostos (reagentes e produtos de reações químicas)	10790	1374 Kb
enzyme	Lista de enzimas, indexada por *EC number*.	4286	6324 Kb
glycan	Informações sobre carboidratos	10017	1340 Kb
reaction	Descrição das reações químicas cadastradas.	5837	1252 Kb
rn_part	Tabela de referências (ligação) entre "reaction" e "compound-glycan"	24245	663 Kb
rn_path	Tabela de referências entre "reaction" e "pathway"	4471	109 Kb
ec_rn	Tabela de referências entre "enzyme" e "reaction"	5365	110 Kb
pathway	Lista de vias metabólicas padrão da KEGG	113	8Kb
genome	Lista de genomas cadastrados	175	230 Kb
genes	Detalhes dos genes dos genomas	603212	1389284 Kb

incluir também a seleção de um organismo. Deste modo, as reações exibidas apresentarão enzimas associadas a elas e estas estarão presentes no genoma do organismo selecionado.

Em termos de sua implementação, ela é composta de formulário web, conforme a Figura 4.2, e de scripts em PHP para execução da consultas à base de dados em MySQL e formatação de resultados.

4.2.1 Interface para a ECPath

A ferramenta ECPath é acessível via internet através do endereço eletrônico: www-.proteome.ib.unicamp.br/tools/ecpath/ (Jul/2005). Através deste endereço, obtém-se o formulário de submissão de dados para a ferramenta, que permite a especificação dos parâmetros necessários para sua execução e, assim, têm-se acesso a suas funcionalidades, conforme descrito abaixo.

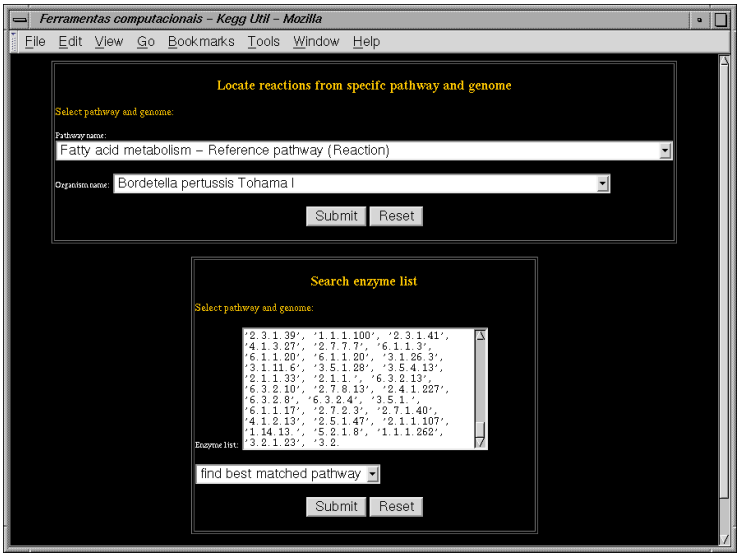


Figura 4.2: Formulário *Web* para submissão de parâmetros na ferramenta ECPath

No caso do uso da ECPath para a localização de conjunto de reações químicas pertencentes a uma via metabólica e também presentes num dado organismo (segunda função apresentada antes), deverão ser preenchidos os 2 primeiros campos exibidos no formulário da Figura 4.2. O primeiro campo exibe uma lista de todas as vias metabólicas de referência cadastradas no banco de dados. Assim, usando este campo, deve-se especificar uma via metabólica de interesse. O segundo campo lista todos os genomas cadastrados e que assim também podem ser selecionados para completar os parâmetros da consulta.

Após a submissão destes dois dados através do botão “Submit”, o usuário obterá como resposta uma página HTML com a listagem das reações químicas localizadas. Conforme mostra a Tabela 4.2, este resultado é composto pelas seguintes informações:

código de identificação da reação química, fórmula da equação química da reação e nome da enzima que realiza a reação.

Tabela 4.2: Reações obtidas para a via metabólica de glicólise e *Xylella fastidiosa* como resultado de execução da ferramenta ECPath.

Reação	Equação da reação	Nome de enzima
R00754	Ethanol + NAD+ <=> Acetaldehyde + NADH + H+	Ethanol:NAD+ oxidoreductase
R00746	Ethanol + NADP+ <=> Acetaldehyde + NADPH	Ethanol:NADP+ oxidoreductase
R00210	Pyruvate + CoA + NADP+ <=> Acetyl-CoA + CO2 + NADPH	Pyruvate:NADP+ 2-oxidoreductase (CoA-acetylating)
R01698	Dihydrolipoamide + NAD+ <=> Lipoamide + NADH	Dihydrolipoamide:NAD+ oxidoreductase
R00014	2-(alpha-Hydroxyethyl)thiamine diphosphate + CO2 <=> Thiamin diphosphate + Pyruvate	2-(alpha-Hydroxyethyl)thiamine diphosphate pyruvate-lyase $(carboxylating)
R01518	2-Phospho-D-glycerate <=> 3-Phospho-D-glycerate	2-Phospho-D-glycerate 2,3-phosphomutase
R01662	3-Phospho-D-glyceroyl phosphate <=> 2,3-Bisphospho-D-glycerate	3-Phospho-D-glyceroyl phosphate 2,3-phosphomutase
R00235	ATP + Acetate + CoA <=> AMP + Pyrophosphate + Acetyl-CoA	Acetate:CoA ligase (AMP-forming)

O outro tipo de consulta disponível na ECPath utiliza os demais campos do formulário da Figura 4.2, principalmente o campo "Enzyme list". Este destina-se ao preenchimento de uma lista de números de identificação de enzima (*EC number*). Os *EC numbers* fornecidos devem estar separados por vírgula ou espaço em branco. A partir desta lista, com o último campo do formulário, pode-se escolher que seja localizado "a via metabólica que melhor se 'encaixa' à lista" ou "o organismo que melhor se 'encaixa' à lista". Tomando como exemplo o primeiro caso (o segundo é análogo), através de buscas à base de dados, cada *EC number* (enzima) fornecida é associada à sua reação química (se existente na base de dados) e cada reação química é associada às vias metabólicas que a contém. Deste modo, quanto maior for o número de enzimas associadas a uma dada via metabólica, melhor seu "encaixe" com a lista de enzimas fornecidas.

Uma lista de enzimas, conforme a que é utilizada na descrição acima, pode representar diferentes situações que as relacionam. Em particular, pode representar um experimento em que houve a expressão de certo número de proteínas (proteômica) e, com a identificação delas, obtém-se uma lista de enzimas referentes a um subconjunto destas proteínas.

Um exemplo de resultado de submissão de dados à ECPath é mostrado na Tabela 4.3. Nela foi empregada o último tipo de consulta descrita acima.

Tabela 4.3: Exemplo de resultado produzido pela *ECPath* para submissão de lista de enzimas originada em experimento proteômico. São mostradas vias metabólicas que apresentaram, ao menos, 2 enzimas relacionadas aos dados de submissão.

Via	Descrição	Número enzimas
61	Biosíntese de ácidos graxos	8
20	Ciclo de Krebs	7
10	Glicólise	6
230	Metabolismo de Purinas	5
720	Reductive carboxylate cycle	4
630	Metabolismo de Glioxilato e dicarboxilato	4
561	Metabolismo de glicerollipideos	3
620	Metabolismo de piruvato	3
710	Fixação de carbono	3
300	Biosíntese de lisina	2
290	Biosíntese de Valina, leucina e isoleucina	2
330	Metabolismo de arginina e prolina	2
30	Via das pentose fosfato	2
51	Metabolismo de Frutose	2
240	Metabolismo de pirimidinas	2
260	Metabolismo de glicina, serina e teonina	2

Capítulo 5

Conclusão

5.1 Discussão geral das ferramentas

Conforme descrito nas seções anteriores, ferramentas computacionais foram desenvolvidas para o apoio à pesquisa em proteômica. Para isso, elas utilizam diversas abordagens de metodologia e manipulam fontes de informação também diversas.

A área de Bioinformática é ramo da ciência recente e que vem crescendo no mundo inteiro, conforme observa-se com o aumento de revistas internacionais especializadas. Neste sentido, novas metodologias e, assim, ferramentas computacionais surgem constantemente. O mesmo podemos observar com relação da ciência, mais experimental, de proteômica. Desta forma, não seria o caso uma expectativa de que as ferramentas aqui apresentadas abrangessem todo o espectro de aplicações possíveis em proteômica. E, naturalmente, um grande espaço de possibilidades resta para desenvolvimentos futuros e criação de novas aplicações que poderão contribuir para o avanço de pesquisas na área.

Apesar desta restrição, podemos observar que as ferramentas aqui apresentadas e discutidas atingem pontos significativos de uma bioinformática voltada para proteômica. Sendo elas:

- Disponibilização de dados e ferramentas via rede (Internet).
- Construção de banco de dados.
- Análise e cálculo de propriedades de seqüências de aminoácidos.
- Espectrometria de massas.
- Utilização de estruturas de dados e conjuntos de dados sobre vias metabólicas.

Nesta linha de pensamento convém então citar algumas áreas tradicionais da bioinformática não abordadas aqui por uma questão de limitação de escopo:

- Análise de genomas.

- Construção e análise de filogenias.
- Algoritmos de busca de seqüências.
- Predição de estruturas secundárias e terciárias de proteínas (bioinformática estrutural)
- Reconhecimento de padrões em expressão gênica.
- Construção de ontologias.

No entanto, o desenvolvimento de atualizações e ampliação do conjunto de funcionalidades destas ferramentas aqui apresentadas são necessários para:

- Melhorar seu uso por parte de pesquisadores envolvidos em projetos proteômica e afins.
- Ampliar o uso em novos projetos.

Para esses fins, é necessária uma melhor integração entre as ferramentas, automatizando procedimentos manuais, e também a criação de recursos que facilitem o acesso às ferramentas.

Especificamente, vemos que as ferramentas de cálculo de pI/MW e criação de Mapa 2D teórico (Seção 2.1 e a ferramenta de estatística N-terminal (Seção 2.2) não fazem uso de proteomas armazenados, requerendo a re-submissão de arquivos FASTA em cálculos e diminuindo a quantidade de dados disponibilizados pelos resultados. Outro fator relacionado a isto é a ausência de referências a fontes de dados externos como o ExPASy (Seção 1.2.2) e outras fontes possíveis.

O Capítulo 3 apresentou recursos para a digestão teórica de proteomas e elaboração de estatísticas sobre identificação de peptídeos usando reagente ICAT (Seção 1.1.4). Em particular, o conjunto de rotinas para digestão teórica não estão disponíveis publicamente e precisam ser executadas manualmente, dificultando a aplicação dos estudos de identificação para novos proteomas. Além disso, uma digestão teórica pode apresentar variação de uma série de parâmetros (dependendo de diversas situações experimentais possíveis) (Wilkins e Williams, 1997), no entanto, este tipo de recurso também não está disponível.

Sobre a ferramenta de vias metabólicas (Capítulo 4), os programas para a conversão direta de dados está limitada ao formato da KEGG. O uso de outros formatos, como do BioCyc, permitiriam uma ampliação dos dados visando obter uma fonte de informações mais completa. Para isso, também, as etapas relacionadas a esta conversão não deveriam ser limitadas a execução manual e, ao contrário, permitir atualização automática para a manutenção dos dados o mais atual possível.

5.2 Sugestões para desenvolvimento futuro

A partir das informações apresentadas sobre proteômica (Capítulo 1) e da discussão sobre ferramentas computacionais desenvolvidas, sobretudo Seção 5.1, apresentamos aqui algumas possibilidades de ampliação e desenvolvimento das ferramentas criadas.

Para cada ferramenta, são descritos as funcionalidades e formas de implementação com o objetivo de:

- Desenvolver modificações e acréscimos de funcionalidades às ferramentas desenvolvidas.
- Promover a integração das ferramentas de modo a formar um "pacote" único de serviços com estas ferramentas, com funcionalidades integradas entre as ferramentas e gerenciamento automático de manutenção de informações.

5.2.1 Banco de dados de proteoma

Com a elaboração de mapas 2D de referência e aplicação de técnicas como *Peptide mass fingerprinting* (Seção 1.1.3), têm-se a necessidade de formas apropriadas para armazenamento dos dados gerados, sendo que, este armazenamento deve possibilitar também o acesso destas informações, publicando-as. Estas características foram contempladas através de um sistema de informações acessível via Internet e que utiliza banco de dados em MySQL e reconhece dados gerado pelo programa de computador normalmente utilizado nas análises de géis 2D (Image-Master).

Esse sistema foi desenvolvido e não está mais detalhado neste texto. Ele armazena informações integradas sobre géis 2D e 1D (imagens e *spots* ou bandas localizados), proteínas identificadas nestes géis e informações genômicas relativas a estas proteínas. Estas informações podem ser acrescentadas através da Internet com o uso de formulários Web específicos, que os tornam públicos, com a opção de especificar algumas restrições de acesso, ver Figura 5.1

O Banco de dados de proteoma poderia ser remodelado para ser o ponto principal de acesso às demais ferramentas.

Para isso, primeiramente, deveria ter a estrutura de tabelas relacionais modificada. Estas modificações compreenderiam:

- Exclusão e modificação de campos das tabelas de genes e genomas visando tornar sua estrutura mais simples para inclusão de proteomas de diversos organismos.
- Inclusão de tabela para projetos específicos. Com isso, diferentes géis armazenados poderão ser armazenados por assunto, ou projeto.
- Inclusão de referências completas entre tabelas de usuários, projetos e géis para implementar maior controle de acesso aos dados armazenados.

As estruturas de dados do banco de dados de proteoma estão implementadas através da ferramenta MySQL e, assim, estas modificações seriam implementadas através deste software.

Além disso, isto exigiria que os *scripts* para inclusão, atualização e consulta a dados através da Internet fossem atualizados para tornar funcionais as mudanças descritas nos dados. E um módulo específico para o gerenciamento de projetos também então seria desenvolvido integrando as funções de: criação de novos projetos, atualização de dados, consulta aos projetos armazenados e atribuição de usuários e géis aos projetos.

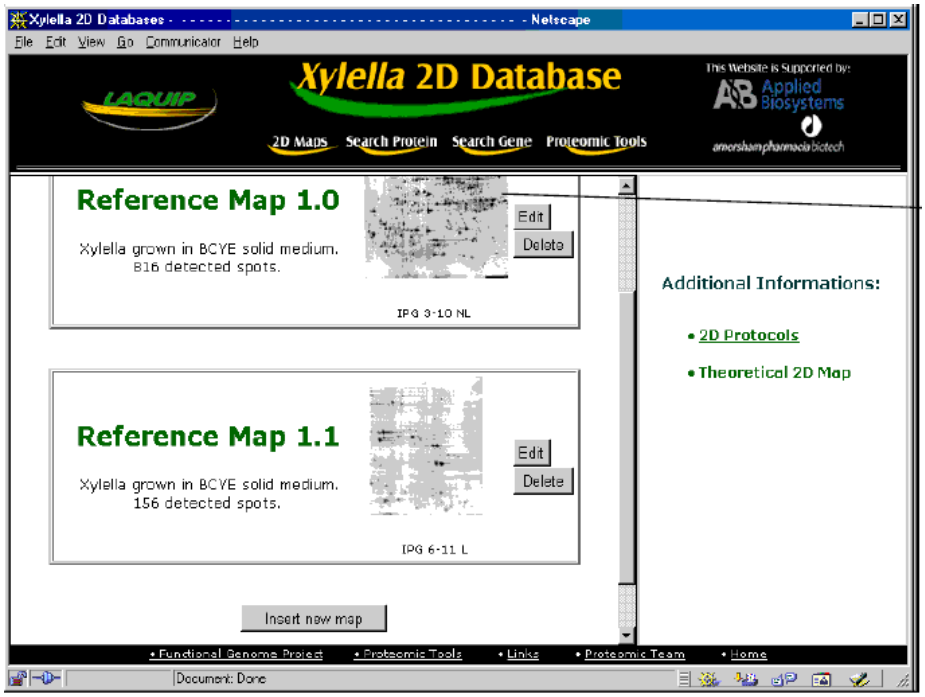


Figura 5.1: Página inicial do banco de dados. Informações armazenadas são utilizadas para gerar páginas HTML como esta exibida. Aqui, mapas de referência 2D soá listados. Sendo que cada um é uma referência para mais informações armazenadas, especificamente sobre as proteínas identificadas em tal mapa

Estes *scripts* seriam implementados através de PHP, em comunicação com a base de dados em MySQL.

As fontes de dados para a inclusão de novos proteomas deveriam incluir:

- A submissão de arquivos FASTA com todas as ORFs previstas para o proteoma a ser incluído.
- A referência direta aos dados sobre genomas provenientes das bases de dados de vias metabólicas.

A submissão dos arquivos FASTA representa uma forma simples de inclusão dos genes/ORFs necessários à identificação de *spots*. Por outro lado, o uso da base de dados de vias metabólicas possibilita a reutiliza de um volume grande de dados já disponíveis, além de construir um ponto importante de integração entre as ferramentas de banco de dados e análise de vias metabólicas.

Outro recurso possível de se implementar é a maior ligação com fontes de dados externas. Dado o caráter dinâmico deste tipo de fonte de dados — devido ao rápido desenvolvimento de novas ferramentas e a possibilidade de mudanças nos endereços eletrônicos para as referências de acesso aos serviços — o mais recomendado é o desenvolvimento de um módulo com *scripts* específicos para cadastramento de serviços e atualização de referências. Deste modo, diversos pontos de *links* externos no sistema de banco de dados poderão ser atualizados de forma centralizada e automatizada. Este módulo seria implementado, a exemplo dos demais, em PHP.

Em relação a *links* às ferramentas internas, a principal forma de implementação seria através do uso de dados armazenados no banco de dados por parte das demais ferramentas, conforme descrito nas próximas seções. Além disso, as páginas que apresentam listas de *spots* identificados de géis poderiam ser modificadas para permitir a submissão destas informações diretamente à ferramenta de vias metabólicas para o estudo de vias relacionadas às proteínas presentes em géis. O Mapa 2D teórico também poderia ser usado de modo que, a partir de um gel experimental em observação, seja construído um gel teórico correspondente a:

- todos os *spots* (correspondentes à proteínas) identificadas no gel, ou;
- todas as proteínas do proteoma (conjunto de ORFs previstas) do organismo em questão.

Em ambos os casos, o mapa teórico poderia ser elaborado na mesma escala que o gel experimental. Para possibilitar a identificação da escala, os *spots* já identificados devem servir de "guias", uma vez que indicam a localização de coordenadas determinadas no mapa experimental.

5.2.2 Mapa 2D teórico e estatísticas N-terminal

A elaboração de mapa 2D teórico e estatísticas N-terminal utiliza, atualmente, como fonte de dados, arquivos FASTA submetidos por usuário. Assim, uma primeira extensão para a ferramenta seria sua modificação para permitir, além das formas já existentes, o uso de conjuntos previamente armazenados de proteomas. Estes dados já

estariam então disponíveis no banco de dados de proteoma, conforme apresentado na Seção 5.2.1.

Na Seção 5.2.1, foi indicada a possibilidade de elaborar-se um mapa teórico na escala obtida de um gel experimental. Esta funcionalidade também poderia estar disponível diretamente a partir da ferramenta de Mapa 2D teórico. Para isso, o formulário de submissão apresentaria uma lista com mapas experimentais incluídos no banco de dados, além também da possibilidade, já implementada, de exibição em mapa com escala linear e escala logarítmica.

Outro recurso possível é a inclusão de referências a dados externos, e internos, nos *spots* dos mapas teóricos, a exemplo dos mapas experimentais. Aqui, "referências internas" aplica-se ao caso de elaboração de mapa teórico a partir de proteomas armazenados.

Estas ferramentas foram implementadas em linguagem Perl e, assim, seus correspondentes *scripts* deveriam ser atualizados para implementar essas atualizações, reformulando também os formulários Web de submissão de informações às ferramentas.

5.2.3 Digestão teórica e identificação de peptídeos

No caso da ferramenta de digestão teórica e identificação de peptídeos via ICAT (Capítulo 3), poderiam ser desenvolvidos formulários para a disponibilidade do uso da ferramenta via Internet (caso específico de digestão, e não da Ferramenta de Busca já apresentada na Seção 3.2.5). Visando a integração das ferramentas, estes formulários, deveriam fazer uso dos proteomas de organismos armazenados no banco de dados de proteoma e também aceitar a entrada de arquivos FASTA que, assim, permitiriam a entrada de proteomas novos.

Os dados resultantes da digestão teórica poderiam permanecer armazenados em forma de banco de dados. Isto seria implementado através do sistema MySQL. Uma vez tendo os dados de digestão armazenados, estes poderão ser aplicados nas ferramentas para localização de peptídeos e também de estatísticas sobre identificação de peptídeos. O acesso destas últimas ferramentas aos bancos de dados de peptídeos deve ficar condicionado ao mesmo sistema de gerenciamento de usuários que o utilizado no banco de dados de proteoma (Seção 5.2.1). Com isso, deverão também ser construídos, via PHP, um sistema de formulários Web para gerenciamento destes bancos de peptídeos.

5.2.4 Vias metabólicas

As ferramentas par análise por vias metabólicas fazem uso, sobretudo, de informações bancos públicos sobre vias metabólicas. Assim, os programas criados para conversão de dados da KEGG poderiam ganhar módulos novos para a realização de atualização automáticas das versões locais destes bancos. Isto permitiria o uso de informações as mais atuais sobre as reações, enzimas e vias metabólicas em estudo. Para a implementação disto, programas em linguagem de programação Java seriam desenvolvidos e realizariam as seguintes etapas:

- *Download* dos dados no formato disponível nos bancos de dados públicos.

- *Parsing* (identificação dos campos de informação) dos arquivos recebidos e simultânea avaliação dos formatos dos dados.
- Inclusão dos dados na base de dados MySQL local.

Atualmente, são utilizados dados do banco de dados KEGG (Seção 1.3.1) para compor a base de vias metabólicas. Esse sistema de atualização automática poderia contemplar também o banco de dados BioCyc (Seção 1.3.2), visando assim obter uma quantidade maior e mais completa de informações.

Uma segunda expansão nas ferramentas de vias metabólicas seria a "construção automática de vias metabólicas". Esta nova ferramenta poderia, a partir de uma lista de enzimas fornecidas, obter vias metabólicas possíveis que as integrem. Aqui, as "vias metabólicas" indicam conjunto de reações químicas que utilizam as enzimas fornecidas (através de números de identificação) ligando-as através de encadeamento $substrato \rightarrow reação \rightarrow produto$, onde a *reação* está ligada a enzima que a realiza. Com o mesmo objetivo de integração entre as ferramentas já exposto, as listas de enzimas utilizadas se originariam das informações sobre proteínas identificadas nos géis armazenados no banco de dados de proteoma. Esta ferramenta seria implementada através de formulários Web, via PHP, para obtenção de parâmetros necessários à execução dos algoritmos de construção de vias e também através de programas em linguagem Perl e Java.

Apêndice A

Produção bibliográfica

A.1 Registros de programas de computador

Universidade Estadual de Campinas. Galembeck, E. & Brum, I.J.B. (2006) *ECPath*. Registro INPI em submissão.

Universidade Estadual de Campinas. Galembeck, E. & Brum, I.J.B. (2001) *Estatística e Busca de N-terminais*. Registro INPI sob nº 39465.

Universidade Estadual de Campinas. Galembeck, E. & Brum, I.J.B. (2000) *Cálculo de pI, MW e eletroforese bi-dimensional teórica*. Registrado no INPI sob nº 00034442.

A.2 Artigo completo publicado em periódico

SMOLKA, Marcus Bustamante; MARTINS, Daniel; WINCK, Flávia Vischi; SANTORO, Carlos Eduardo; CASTELLARI, Rafael Ramos; FERRARI, Fernanda; BRUM, I. J.; GALEMBECK, Eduardo; COLETTA FILHO, Helvécio Della; MACHADO, Marcos Antônio; MARANGONI, Sérgio; NOVELLO, José Camillo. Proteome analysis of the plant pathogen Xylella fastidiosa reveals major cellular and extracellular proteins and a peculiar codon bias distribution. *Proteomics (Weinheim. Print)*, NY USA, v. 3, p. 224-237, 2003.

A.3 Resumos publicados em anais de eventos

Milani, R.; Brum, I. J. B.; Martins, A. R.; Galembeck, E. *Development of Computational Tools for a Dynamic Association between Genomic and Proteomic Data and Metabolic Pathways*. Em: ISMB - 14th Annual International Conference On Intelligent Systems For Molecular Biology, 2006, Fortaleza-CE. ISMB2006 - Conference Proceedings, 2006.

Brum, I. J. B.; Galembeck, E. *Evaluation of computational tool for peptide identification based on multiple database constrains*. Em: Swiss-Prot 20 Years: In-Silico Analysis of Proteins - Celebrating the 20th Anniversary of Swiss-Prot, 2006, Fortaleza-CE. Program and Abstract Book.

Martins, A. R.; Brum, I. J.; Kitajima, J. P. F. W.; Galembeck, E. *Integrating Sequence Data in Metabolic Map Model*. Em: XXXIV Reunião Anual da SBBq, 2005, Águas de Lindóia-SP. CD de Programas e Resumos, 2005.

Brum, I. J.; Martins, A. R.; Galembeck, E. *New features in ProteomeDB: a dabase integrated system for Proteomics*. Em: X-meeting - 1ª Conferência Internacional da AB3C, 2005, Caxambu-MG. Livro de Resumos do X-Meeting - 1ª Conferência Internacional da AB3C, 2005.

BRUM, I. J.; MARTINS, Anderson Rodrigues; GALEMBECK, Eduardo. New features in ProteomeDB: a dabase integrated system for Proteomics. Em: X-meeting - 1ª Conferência Internacional da AB3C, 2005, Caxambu-MG. Livro de Resumos do X-Meeting - 1ª Conferência Internacional da AB3C, 2005.

OLIVEIRA, Lígia P; URBANO, Roberta R; BRUM, I. J.; NOVELLO, José Camillo; GALEMBECK, Eduardo. A Theoretical Method to Evaluate the Accuracy of Peptides Identification Based on their Mas and Isoeletric Point. Em: XXXIII Reunião Anual da SBBq, 2004, Caxambu-MG. CD de Programas e Resumos, 2004.

BRUM, I. J.; GALEMBECK, Eduardo. Uma Ferramenta para Estudo de Função Biológica de Genes Correlacionando Informações de Genoma, Transcriptoma e Proteoma. Em: XII Congresso Interno de Iniciação Científica da Unicamp, 2004, Campinas. Carderno de Resumos do XII Congresso Interno de Iniciação Científica. Campinas : Unicamp, 2004. p. 75.

BRUM, I. J.; GALEMBECK, Eduardo. A Web Based Tool for Metabolic Pathways Analysis. Em: International Conference on Bioinformatics and Computational Biology, 2004, Angra dos Reis-RJ/Brasil, 2004.

BRUM, I. J.; MARTINS, Daniel; SMOLKA, Marcus B; NOVELLO, José Camillo; GALEMBECK, Eduardo. Database Integrated System for Proteomics Analysis. Em: XXXII Reunião Anual da SBBq, 2003, Caxambu-MG. Caderno de Resumo da XXXII Reunião Anual da da SBBq. São Paulo, 2003. p. 261.

BRUM, I. J.; SMOLKA, Marcus B; NOVELLO, José Camillo; GALEMBECK, Eduardo. Theoretical Avaliation of a New Protein Identification Method Based on Multiple Database Constrains. Em: XXXI Reunião Anual da SBBq, 2002, Caxambu-MG. Caderno de Resumos da XXXI Reunião Anual da SBBq, 2002. p. 262.

BRUM, I. J.; GALEMBECK, Eduardo; SMOLKA, Marcus B; NOVELLO, José Camillo. Avaliação Teórica de um Novo Método de Identificação de Proteínas Baseado em Múltiplas Restrições a Base de Dados. Em: X Congresso Interno de

Iniciação Científica da Unicamp, 2002, Campinas-SP. Carderno de Resumos do X Congresso Interno de Iniciação Científica. Campinas : Unicamp, 2002. p. 56.

BRUM, I. J.; SMOLKA, Marcus B; NOVELLO, José Camillo; GALEMBECK, Eduardo. Calculation of pI, MW and Theoretical 2D Map: a Computational Tool for Proteome Analysis. Em: XXX Reunião Anual da SBBq, 2001, Caxambu-MG. Caderno de Resumo da XXX Reunião Anual da SBBq. São Paulo, 2001. p. 222-222.

BRUM, I. J.; SMOLKA, Marcus B; MARTINS, Daniel; NOVELLO, José Camillo; GALEMBECK, Eduardo. Development of Computational Tools for the Xylella fastidiosa Proteome Project. Em: I Simpósio Genoma Funcional da Xylella fastidiosa, 2001, Serra Negra-SP. Caderno de Resumos do I Simpósio Genoma Funcional da Xylella fastidiosa, 2001. p. 77.

MARTINS, Daniel; SMOLKA, Marcus B; WINCK, Flávia Vischi; SANTORO, Carlos e; BRUM, I. J.; GALEMBECK, Eduardo; MARANGONI, Sérgio; NOVELLO, José Camillo. Database Constrain Tool for Protein Identification Based on pI and MW Prediction from Genome Sequence. Em: XXX Reunião Anual da SBBq, 2001, Caxambu-MG. Caderno de Resumos da XXX Reunião da SBBq. São Paulo, 2001. p. 166.

SMOLKA, Marcus B; WINCK, Flávia Vischi; MARTINS, Daniel; SANTORO, Carlos e; BRUM, I. J.; GALEMBECK, Eduardo; C FILHO, H. D.; MACHADO, M. A.; LEMOS, E. G. M.; TOYAMA, M. H.; MARANGONI, Sérgio; NOVELLO, José Camillo. The Xylella fastidiosa Proteome Database. Em: I Simpósio Genoma Funcional da Xylella fastidiosa, 2001, Serra Negra-SP. Caderno de Resumos do I Simpósio Genoma Funcional da Xylella fastidiosa, 2001. p. 29.

BRUM, I. J.; SMOLKA, Marcus B; MARTINS, Daniel; SANTORO, Carlos e; NOVELLO, José Camillo; GALEMBECK, Eduardo. Ferramenta Computacional para Comparação Visual de Proteoma. Em: IX Congresso Interno de Iniciação Cinetífica da UNICAMP, 2001, Campinas. Caderno de Resumos do IX Congresso Interno de Iniciação Científica. Campinas: Unicamp, 2001. p. 52.

BRUM, I. J.; SMOLKA, Marcus B; NOVELLO, José Camillo. Software para cálculo de ponto isoelétrico e peso molecular teóricos como ferramenta para análise de proteoma. Em: XV Reunião Anual da FeSBE, 2000, Caxambu-MG, 2000.

BRUM, I. J.; SMOLKA, Marcus B; NOVELLO, José Camillo; GALEMBECK, Eduardo. Desenvolvimento de Algoritmos para Cálculo de Ponto Isoelétrico e Peso Molecular para Análise de Proteoma. Em: VIII Congresso Interno de Iniciação Científica da UNICAMP, 2000, Campinas-SP. Caderno de Resumos do VIII Congresso de Iniciação Científica. Campinas: Unicamp, 2000. p. 36.

Referências Bibliográficas

Biosystems, A. (2001) *ICAT*^{TM} *Kit for Protein labeling*, p. 2.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. e Schneider, M. (2003) The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Cahill, D. J., Nordhoff, E., O'Brien, J., Klose, J., Eickhoff, H. e Lehrack, H. (2001) Bridging genomics and proteomics. Em Pennington, S. R. e Dunn, M. J. (eds.), *Proteomics: From Protein Sequence to Function*. BIOS Scientific Publishers, Oxford.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. e Bairoch, A. (2003) Expasy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. e Bairoch, A. (2005) Protein identification and analysis tools on the expasy server. Em Walker, J. M. (ed.), *The Proteomics Protocols Handbook*, pp. 571–607. Humana Press Inc., Totowa, NJ.

Grosu, P., Townsend, J. P., Hartl, D. L. e Cavalieri, D. (2002) Pathway processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res.*, **12**, 1121–1126.

Gygi, S. P. e Aebersold, E. (2000) Using mass spectrometry for quantitative proteomics. Em *Proteomics a Current Trends Supplement*, pp. 32–37. NA.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H. e Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, **17**, 994–9.

Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C. e Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in proteins sequence databases. *Proceedings of the National Academy of Sciences of USA*, **90**, 5011–5015.

Herrman, E. (1997) *Aprenda em 1 semana programação CGI com PERL 5*. Campus, Rio de Janeiro, Brasil.

Hoogland, C., Mostaguir, K., Sanchez, J.-C., Hochstrasser, D. F. e Appel, R. D. (2004) Swiss-2dpage, ten years later. *Proteomics*, **4**, 2352–2356.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R. e Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

James, P., Quadroni, M., Carafoli, E. e Gonnet, G. (1993) Protein identification by mass profile fingerprinting. *Biochemical and Biophisical Research Communications*, **195**, 58–64.

Kahaner, D., Moler, C. e Nash, S. (1989) *Numerical Methods and Software*, pp. 240–242. Prentice Hall PTR.

Kanehisa, M. (1997) A database for post-genome analysis. *Trends Genet*, **13**, 375–376.

Kanehisa, M. (2000) *Post-genome Informatics*. Oxford University Press.

Kanehisa, M. e Goto, S. (2000) Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kanehisa, M., Goto, S., Kawashima, S. e Nakaya, A. (2002) The kegg databases at genomenet. *Nucleic Acids Res.*, **30**, 42–46.

Karp, P. D., Riley, M., Paley, S. M. e Pellegrini-Toole, A. (2002) The metacyc database. *Nucleic Acids Res.*, **30**, 59–61.

Lahm, H. W. e Langen, H. (2000) Mass spectrometry: a tool for the identification of proteins separated by gels. *Electrophoresis*, **21**, 2105–14.

Lemay, L. (1999) *Teach Yourself Web Publishing with HTML*. Sans Net.

Lindroos, H. e Andersson, S. G. E. (2002) Visualizing metabolic pathways: Comparative genomics and expression analysis. Em *Proceedings of the IEEE*, volume 90, pp. 1793–1802.

Michal, G. (1998) On representation of metabolic pathways. *Biosystems*, **47**, 1–7.

Ogata, H., Goto, S., Fujibuchi, W. e Kanehisa, M. (1998) Computation with the kegg pathway database. *Biosystems*, **47**, 119–128.

Pappin, D. J. C., Hojrup, P. e Bleasby, A. J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, **3**, 327–332.

Patrickios, C. S. e Yasaki, E. (1995) Polypeptide amino acid composition and isoelectric point: comparison between experiment and theory. *Analytical Biochemistry*, **231**, 82–91.

Patterson, S. D., Aebersold, R. e Goodlett, D. R. (2001) Mass spectrometry-based methods for protein identification and phosphorylation site analysis. Em Pennington, S. R. e Dunn, M. J. (eds.), *Proteomics: From Protein Sequence to Function*. BIOS Scientific Publishers, Oxford.

Patton, W. F. (2001) Detection proteins in polyacrylamide gels and on electroblot membranes. Em Pennington, S. R. e Dunn, M. J. (eds.), *Proteomics: From Protein Sequence to Function*. BIOS Scientific Publishers, Oxford.

Pleissner, K. P., Oswald, H. e Wegner, S. (2001) Mass spectrometry-based methods for protein identification and phosphorylation site analysis. Em Pennington, S. R. e Dunn, M. J. (eds.), *Proteomics: From Protein Sequence to Function*. BIOS Scientific Publishers, Oxford.

Scraton, R. E. e Arnold, E. (1986) *Basic Numeric Methods: An Introdution to Numerical Mathematics on a Microcomputer*, p. 25. Publisers Ltd.

Smolka, M. B., Zhou, H., Purkayastha, S. e Aebersold, R. (2001) Optimization of the isotope-coded affinity tag-labeling procedure for quantitative proteome analysis. *Anal Biochem*, **297**, 25–31.

Urquhart, B. L., Cordwell, S. J. e Humphery-Smith, I. (1998) Comparison of predicted and observed properties of proteins encoded in the genome of mycobacterium tuberculosis h37rv. *Biochem Biophys Res Commun*, **253**, 70–9.

Westermeier, R. (1997) *Electrophoresis in Practice*. VCH, Weinheim, Germany.

Wilkins, K. L. e Hochstrasser, D. F. (1997) *Proteome Research: New Frontiers in Functional Genomics.*, pp. 1–11. Springer-verlag, Berlin Heidelberg, Germany.

Wilkins, M. R., Gasteiger, E., Tonella, L., Ou, K., Tyler, M., Sanchez, J. C., Gooley, A. A., Walsh, B. J., Bairoch, A., Appel, R. D., Williams, K. L. e Hochstrasser, D. F. (1998) Protein identification with n and c-terminal sequence tags in proteome projects. *J Mol Biol*, **278**, 599–608.

Wilkins, M. R. e Gooley, A. A. (1997) *Proteome Research: New Frontiers in Functional Genomics.*, pp. 35–61. Springer-verlag, Berlin Heidelberg, Germany.

Wilkins, M. R. e Williams, K. L. (1997) Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: A theoretical evaluation. *Journal of Theoretical Biology*, pp. 7–15.