COPPE/UFRJ

UM AMBIENTE COMPUTACIONAL PARA PROTEÔMICA

Paulo Costa Carvalho

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Valmir Carneiro Barbosa

Rio de Janeiro
Março de 2010

UM AMBIENTE COMPUTACIONAL PARA PROTEÔMICA

Paulo Costa Carvalho

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

Examinada por:

____
Prof. Valmir Carneiro Barbosa, Ph.D.

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

____
Dr. Wim Maurits Sylvain Degrave, Ph.D.

____
Dr. Jonas Enrique Aguilar Perales, D.Sc.

____
Dr. Dário Eluan Kalume, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

MARÇO DE 2010

Carvalho, Paulo Costa

Ambiente Computacional para Proteômica / Paulo Costa Carvalho. – Rio de Janeiro: UFRJ/COPPE, 2010.

XI, 91 p.: il.; 29,7 cm.

Orientador: Valmir Carneiro Barbosa

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2010.

Referencias Bibliográficas: p. 42-44.

1. Proteômica. 2. Reconhecimento de Padrões. 3. Espectrometria de Massa. I. Barbosa, Valmir Carneiro. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Titulo.

Dedico esta tese aos meus pais,

José Francisco de Oliveira Carvalho
e Maria da Gloria da Costa Carvalho

“Think of all the men who never knew the answers, think of all those who never even cared, still there are some who asked why, who wanted to know, who dared to try.”

Rod Mckuen, “Here he comes again”

Agradecimentos

Agradeço ao Prof. Valmir Carneiro Barbosa, mais que um orientador, um grande amigo que embarcou junto em uma jornada de desafios multi-diciplicares; ao amigo Elias Bareinboim por toda a contribuição; ao Prof. John R Yates III pela amizade, e por ter me recebido em seu laboratório no Scripps Research Institute; aos colegas de lá, em particular Daniel Cociorva, Tao Xu, Xuemei Han, e Sherry Niessen pelo aprendizado. Agradeço também aos Prof. Gilberto Barbosa Domont e aos colegas Fabio e Carlos pela troca de idéias; ao apoio dos amigos da Fiocruz, em especial do Dr. Jonas Perales e do Dr. Wim Degrave e colegas de seus respectivos laboratórios; à minha esposa Juliana de Saldanha da Gama Fischer Carvalho por ter embarcado junto comigo neste desafio, pelo trabalho em conjunto, e pela nossa linda filha Giulia nascida durante a elaboração desta tese; ao grande amigo Royler Gracie, especialmente pelos treinos em San Diego que ajudaram manter "*Mens sana in corpore sano*".

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

UM AMBIENTE COMPUTACIONAL PARA PROTEÔMICA

Paulo Costa Carvalho

Março/2010

Orientador: Valmir Carneiro Babosa

Programa: Engenharia de Sistemas e Computação

O presente trabalho introduz novas metodologias computacionais para analisar dados de proteômica *shotgun*. A primeira, o *PatternLab for proteomics*, cria ambiente computacional capaz de apontar proteínas diferencialmente expressas entre perfis protéicos, analisa dados proteômicos sob a luz do *Gene Ontology*, aponta tendências em experimentos temporais, e gera facilmente diagramas Venn com áreas proporcionais. A segunda contribuição, intitulada *Charge Prediction Machine* (CPM), infere a carga de íons precursores com base no espectro de massa de baixa resolução em tandem (1000) obtido por dissociação de transferência de elétron; o conhecimento da carga é necessário para a identificação protéica. O CPM utiliza abordagem inovadora inspirada no discriminante bayesiano; comparativamente, enquanto nossa abordagem acertou 98% da carga dos precursores em um gabarito, a única metodologia existente (Charger, Thermo Fisher, San José – CA) acertou 86%. A terceira contribuição, intitulada YADA, introduz algoritmo para deconvolução de espectros de alta resolução e acurácia (<50 ppm). Quando comparado ao comercialmente disponível (o Xtract da Thermo Fisher, San Jose, CA), YADA mostrou-se 700% mais rápido e aumentou em 20% o número de peptídeos identificados. Em seguida, introduz-se uma nova metodologia experimental / computacional para aquisição de dados de proteômica *shotgun* intitulado *Extended Data Independent Analysis* (XDIA). Quando testado em um lisado de levedura, aumentaram-se em 250% os espectros identificados e 30% no número de peptídeos únicos quando comparado à metodologia estado da arte e largamente adotada: *data-dependent analysis*.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

A COMPUTATIONAL ENVIORNMENT FOR PROTEOMICS

Paulo Costa Carvalho

March/2010

Advisor: Valmir Carneiro Barbosa

Department: Systems Engineering and Computer Science

This thesis contributes with methodologies for analyzing shotgun proteomic data. Firstly, we introduce PatternLab for proteomics, a computational environment that can pinpoint differentially expressed proteins when analyzing data from complex peptide mixtures, leverage the Gene Ontology to aid in experimental interpretation, discriminate trends in time-course experiments, and easily generate approximately area proportional Venn diagrams. Secondly, we introduce Charge Prediction Machine (CPM). CPM infers precursor ion charge based on its low resolution tandem mass spectrum (~1000) acquired using electron transfer dissociation; knowledge of the precursor charge is necessary for protein identification. CPM relies on a new approach inspired in the Bayesian discriminant function; it correctly classified 98% of the precursor charges in a test database while the only competing methodology (Charger, Thermo Fisher, San Jose – CA) correctly classified 86%. Thirdly, we introduce YADA, a new algorithm for deconvolution of high-resolution, high-accuracy mass spectra (<50 ppm). When compared to the commercially available solution (Xtract, Thermo Fisher, San Jose, CA), YADA identified 20% more unique peptides and was 700% faster. Lastly, we introduce a new experimental / computational approach for shotgun data acquisition called Extended Data Independent Analysis (XDIA). When tested on a yeast lysate, the number of identified spectra and unique peptides increased by 250% and 30%, respectively, as when compared to the state-of-the-art and widely adopted data-dependent analysis.

Sumário

Lista de abreviaturas, símbolos e unidades

AAPVD	Diagrama de *Venn* de área proporcional
A 172	Linhagem celular de glioblastoma multiforme
AG	Algoritmo Genético
AP	Álcool perílico
CAD ou CID	Dissociação assistida por colisão
CPM	*Charge prediction machine*
DDA	*Data dependent analysis*
DIA	*Data independent analysis*
ESI	Ionização por spray de elétrons do inglês *Electron Spray Ionization*
ETD	Dissociação por transferência de elétron
GBM	Glioblastoma Multiforme
Da	Dalton
LC/LC	Cromatografia líquida em duplo estágio
LC/LC/MS/MS	Cromatografia líquida em duplo estágio acoplada à espectrometria de massa em tandem
LTQ	*Linear Trap Quadrupolo*
MS	Espectro de massa
MS1	Espectro do perfil de massas
MS2 ou MS/MS	Espectrômetro do perfil de massas de íons dissociados
m/z	Razão massa/carga
MudPIT	*Multidimensional Protein Identification Technology*
ppm	Parte por milhão
RF	Radiofreqüência
RP	Fase Reversa; do inglês *Reverse Phase*
SCX	Troca catiônica forte; do inglês *Strong Cation Exchange*
SVM	Máquina de vetor de suporte
XDIA	*Extended Data Independent Analysis*

1. Introdução

1.1. O desafio

A evidenciação de padrões diferenciais na expressão de proteínas possibilita melhor compreensão de diversos fenômenos biológicos. In vitro, destacam-se aplicabilidade no estudo das respostas de células em cultura a fármacos, alterações do meio ou estímulos externos; *in vivo,* a caracterização de biomarcadores, na medicina personalizada para auxiliar diagnóstico precoce, e prognóstico de patologias.

A proteômica introduziu metodologias capazes de retratar diferenciais entre perfis complexos de proteínas e, assim, permitiu o estudo em larga escala de milhares de proteínas de uma só vez. Dentre estas metodologias, a proteômica *shotgun* é considerada estado-da-arte; ela utiliza cromatografia líquida acoplada a espectrometria de massa. Atualmente, é inconcebível analisar qualquer processo biológico ou elaborar a criação de novos painéis de proteínas marcadoras para diagnóstico (biomarcadores) sem considerar o uso da proteômica. Porém, os dados gerados por esta tecnologia são numerosos e de interpretação difícil; isto implica a necessidade do desenvolvimento de algoritmos especializados.

As primeiras referências sobre diagnóstico de enfermidades por biomarcadores datam de 560 anos a.C. quando Hipócrates propôs que as modificações em líquidos orgânicos, dentre eles o sangue, estavam associadas a patologias. A partir desta época, as doenças foram mais bem caracterizadas pela busca de alterações em nossos "elementos constitutivos". Ao final do século XX e início do XXI, foram iniciadas pesquisas para definir a nossa "identidade molecular normal" e as suas modificações devidas a doenças, visando aumentar o poder do diagnóstico. Cunhou-se o termo "biomarcador" para qualquer molécula ou característica biológica que possa ser detectada e medida revelando os processos biológicos normais, patogênicos ou a resposta farmacológica após intervenção terapêutica. Desta forma, o biomarcador deve indicar alterações do estado fisiológico normal para o patológico ou de mudanças no ambiente corporal interno ou no meio ambiente. No campo molecular as pesquisas focalizaram a atenção sobre nossos genes, RNAs, proteínas e moléculas ligadas ao metabolismo. Consequentemente, o conjunto destas moléculas foi denominado respectivamente: genoma[1], transcriptoma[2], proteoma[3], e metaboloma para

[1] Todo o material genético contido na célula.

relacionar os metabólitos presentes em determinadas situações. Estas definições podem ser usadas em relação a um organismo, um tecido, fluido corporal ou célula.

No ano de 2001 foi concluída e publicada a primeira versão da sequência do genoma humano (1). Em paralelo, inúmeros laboratórios realizavam a busca para identificar as proteínas expressas no organismo, tecido, fluido corporal ou célula humana e atribuir a essas moléculas uma função. A pretensão de conhecer os genes, mRNAs, metabólitos e proteínas da célula, em determinadas condições, constitui um dos maiores desafios das ciências biológicas.

Sabe-se atualmente que a expressão gênica pode ser transcricional ou pós-transcricionalmente regulada. Durante as últimas décadas a análise quantitativa de transcritos gênicos mostrou não haver proporcionalidade direta entre o nível destes e o nível da proteína presente na célula em resposta a um estímulo específico ou estado patológico (2). Isto se deve à existência de numerosos mecanismos de controle da expressão gênica que operam durante a tradução (3-6). Em outras palavras, o quadro transcricional qualitativo e quantitativo da célula pode não corresponder àquele encontrado para proteínas. Muitas espécies de mRNAs são sintetizadas em resposta a perturbações específicas e possuem vida média curta; outras populações podem estar presentes, porém podem não estar sendo traduzidas (7). Outros problemas de mesma natureza decorrem das modificações pós-transcricionais experimentadas pelos RNAs (*splicing*) e pós-traducionais sofridas por muitas proteínas. Na comparação direta entre o genoma e o proteoma a situação é ainda mais complexa, considerando os mecanismos de *splicing* alternativo[4] que aumentam, em muito, a capacidade de codificação de um determinado gene (8). Consequentemente, as pesquisas foram direcionadas para conhecer o proteoma celular na forma padrão ou em situações onde as funções celulares foram perturbadas ou encontrem-se em situações patológicas específicas.

Assim como o conhecimento do genoma se deve, praticamente, ao advento da engenharia genética e as metodologias de sequenciamento de DNA, o conhecimento do proteoma foi alavancado, principalmente, graças ao aparecimento da eletroforese bidimensional (2D) de alta

[2] Todos os transcritos de RNA.

[3] O conjunto de proteínas expressas.

[4] Método pelo qual íntrons são removidos, e exons são unidos para formar o RNA mensageiro.

resolução, à cromatografia líquida e às técnicas de espectrometria de massa capazes de ionizar e analisar uma biomolécula sem degradá-la (9,10).

Nossa individualidade molecular implica o uso de múltiplos biomarcadores na construção de modelos probabilísticos para diagnósticos diferenciados e personalizados. A identificação de painéis de biomarcadores desafia o campo da proteômica exigindo cada vez mais sensibilidade e poder de quantificação que os das técnicas existentes (eletroforese em gel, cromatografia e espectrometria de massa[5]). Esta problemática desafia também a ciência de inteligência artificial no setor de reconhecimento de padrões. Outros fatores limitantes são: o custo dos equipamentos e reagentes, o elevado número de parâmetros por amostra[6], a grande variabilidade entre amostras de mesma classe, as limitações na reprodutibilidade das técnicas proteômicas para detecção e quantificação simultânea de milhares de proteínas e a falta de conhecimento de funções de densidade de probabilidade que descrevam adequadamente a variação do nível de expressão de proteínas para o caso em estudo.

A identificação dos componentes de um sistema biológico, assim como a caracterização de padrões diferenciais se fazem necessárias para uma interpretação holística do sistema biológico. Contudo, a elevada quantidade de informação gerada pelas técnicas "ômicas[7]", quando aplicadas a analitos complexos (*i.e.* lisados celulares, soro e tecidos), "coloca em cheque" inclusive os melhores especialistas da área. Um dos objetivos da proteômica é caracterizar os estados de um sistema biológico através de alterações no perfil de expressão de proteínas. Inúmeros trabalhos comparam perfis proteômicos de amostras obtidas de pacientes com os de indivíduos do grupo de controle. Uma atenção especial é dada ao diagnóstico precoce de neoplasias, no qual a técnica parte da premissa que biomoléculas podem se originar em microambientes do tumor-hospedeiro (biomarcadores específicos) ou na resposta do sistema imunológico (marcadores inespecíficos) à patologia. O fato estabelece um novo paradigma, no qual a tecnologia dos "omas"[8] aliada a

[5] O espectrômetro de massa é um instrumento analítico capaz de converter moléculas neutras em íons na forma gasosa e separá-las de acordo com a sua razão massa/carga (*m/z*).

[6] Em uma única amostral de soro, certas técnicas proteômicas podem identificar e quantificar mais de 1.000 proteínas.

[7] Proteômica, genômica, etc.

[8] Genoma, proteoma, metaboloma, interatoma, etc.

metodologias de reconhecimento de padrões poderá trazer um impacto direto na clínica médica.

A necessidade de aumentar o número de proteínas identificadas em misturas complexas levou ao desenvolvimento da tecnologia multidimensional de identificação de proteínas; do inglês, *Multi-dimensional Protein Identification Technology* (MudPIT) Resumidamente, o MudPIT utiliza coluna de troca iônica seguida de coluna de fase reversa diretamente acoplada ao espectrômetro de massa. A cromatografia bi-dimensional realiza-se aplicando na eluição a função degrau de aumento de concentração salina liberando pacotes de peptídeos da coluna de troca iônica para a coluna da fase reversa (RP). Cada eluato obtido da coluna de troca iônica é posteriormente submetido a gradiente hidrofóbico na coluna RP e os peptídeos identificados por MS/MS. A técnica usa, geralmente, doze degraus de concentração salina produzindo doze corridas cromatográficas de RP. Devido às condições empregadas na cromatografia líquida 2D (LC/LC), na mesma amostra, peptídeos de mesma classe podem não eluir necessariamente no mesmo passo salino. Isto torna o alinhamento cromatográfico praticamente impossível, tornando inviável a aplicação das técnicas existentes de quantificação a dados de LC/LC/MS/MS.

Para contornar este problema, Liu *et. al.* (11) desenvolveu uma metodologia que permite a quantificação relativa de proteínas em dados de MudPIT, correlacionando as proteínas a seus números de identificações MS/MS (ou contagem espectral); a técnica foi demonstrada através da adição de proteínas marcadoras a amostras de lisados celulares. Contudo a técnica ainda precisava mostrar-se suficientemente sensível para utilização em estudos de proteômica diferencial.

A introdução de espectrômetros de alta resolução (*e.g.*, Orbitrap XL) e das metodologias para dissociação de biomoléculas (*e.g.*, dissociação por transferência de elétron (ETD) (12)) trouxeram também novos desafios tecnológicos. Dentre eles, a necessidade de aperfeiçoamento de ferramentas de busca e novos algoritmos que permitam aproveitamento máximo de dados de alta resolução na identificação proteica. Os algoritmos pré-existentes não estavam aptos a analisar dados de ETD. O acoplamento "ETD-algoritmo" é fundamental na análise de grandes moléculas proteicas (maiores que as obtidas por digestão tríptica) e no estudo de modificações pós-traducionais. Esse cenário provocou o início dos trabalhos desta tese.

1.2. Metodologia de espectrometria de massa usada nesta tese

1.2.1 Espectrometria de massa para proteômica shotgun

O espectrômetro de massa é um instrumento analítico capaz de converter moléculas neutras em íons na forma gasosa e separá-las de acordo com a sua razão/massa carga (*m/z*). Basicamente, este instrumento é composto por uma fonte ionizante, analisador(es) e detector(es), conforme esquematizado na Figura 1. Durante a operação, primeiramente, o analito é ionizado na fonte e acelerado por campo elétrico, e transforma a sua energia potencial elétrica em energia cinética. Na segunda etapa, o(s) analisador(es) separa(m) íons de acordo com sua razão *m/z* para obtenção do perfil de massas (MS1), ou selecionam íons com *m/z* previamente escolhidos e os submetem à fragmentação em um processo denominado "espectrometria de massa em tandem" (MS2 ou MS/MS) (descrito adiante). O detector juntamente com o computador compõem a etapa final do processo, registrando e amplificando a corrente iônica oriunda do analito e representando o sinal do espectro de massa a ser interpretado pelo operador ou por programas de bioinformática. O espectro de massas é representado graficamente, empregando a abscissa para quantificar a intensidade da corrente iônica e a ordenada evidenciando a razão massa / carga dos íons[9].

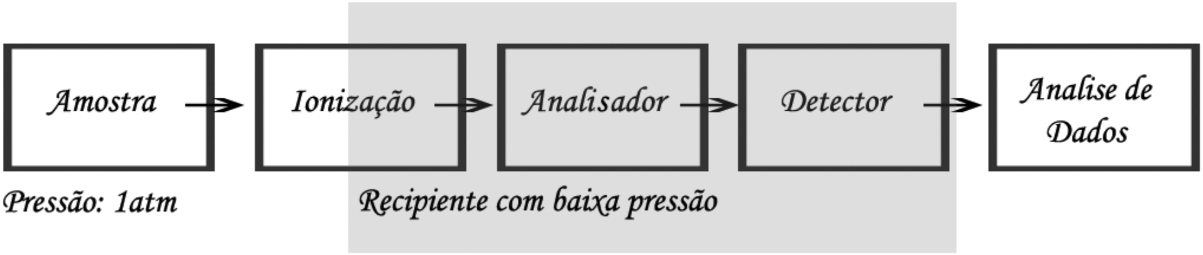


Figura 1 - Esquema simplificado de um espectrômetro de massa

Atualmente existem inúmeros modelos de espectrômetro de massa, cada um otimizado para uma particularidade. As secções subsequentes introduzem o Orbitrap XL (Thermo, San José, Califórnia) (figura 2), que é considerado o estado-da-arte, e foi o equipamento utilizado nos

[9] Nos estudos de proteômica, "os íons" referem-se, quase sempre, aos peptídeos ionizados.

experimentos descritos nesta tese. Os algoritmos aqui desenvolvidos podem, em geral, ser estendidos para outros espectrômetros capazes de analisar dados de proteômica *shotgun*.

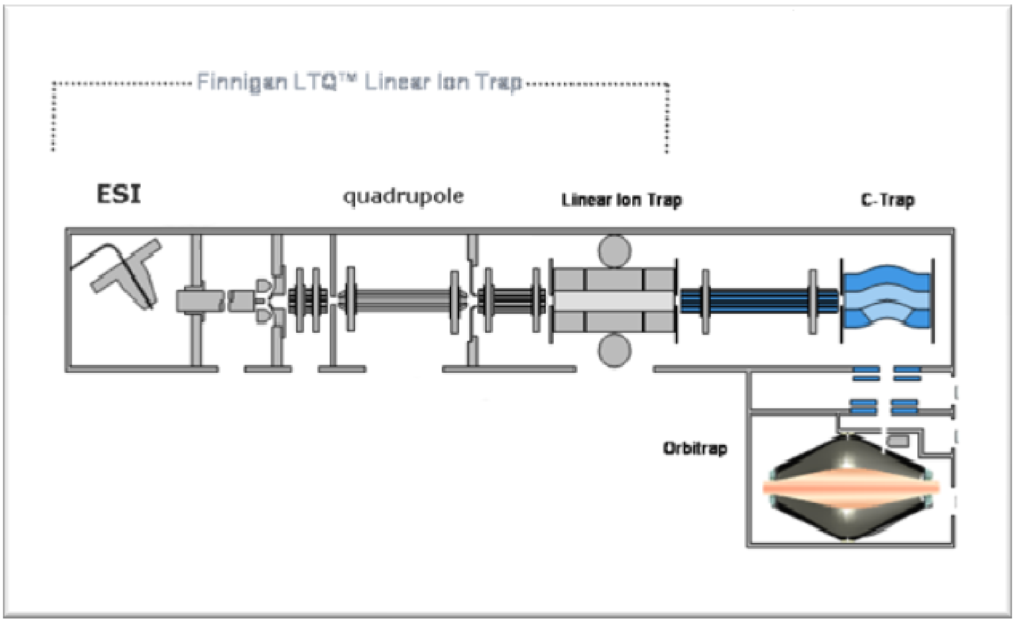


Figura 2 - Esquema de espectrômetro de massa com fonte ionizante *electrospray* e analisadores do tipo quadrupolo/armadilha de íons linear e Orbitrap (figura obtida do manual do espectrômetro, Thermo Fisher, San José CA).

O Orbitrap XL contém fonte ionizante *electrospray* (ESI), analisadores do tipo “armadilha de íons” (*linear-trap* ou LTQ) e Orbitrap. O Orbitrap XL é capaz de realizar ensaios de espectrometria de massa em *tandem*. A proteômica *shotgun* utiliza a espectrometria em tandem na identificação de peptídeos e proteínas. Inicialmente, o espectro de massas das moléculas injetadas no espectrômetro é gerado (MS1). Subsequentemente, a razão/massa carga (*m/z*) correspondente aos “peptídeos íons” que se deseja identificar são obtidos do MS1. Em seguida, o espectrômetro isolará, separadamente, os íons com *m/z* de interesse, e os desassociará. Espera-se que a informação do *m/z* do “íon parental” e do respectivo espectro de massa dos “íons filhos” (espectro de massa em *tandem* obtido da dissociação do íon parental) permita identificar a molécula. A figura 3 mostra um espectro de massa em tandem onde foi possível elucidar a sequência de aminoácidos a partir dos fragmentos da molécula precursora.

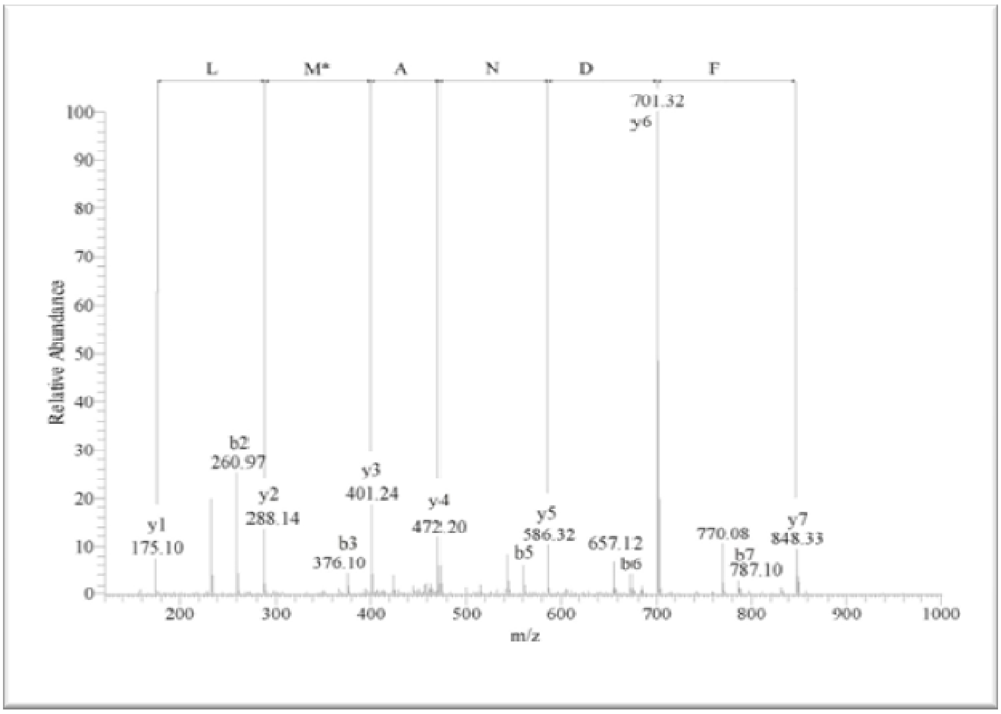


Figura 3 - A sequência aminoacídica do peptídeo FDNAM*L é elucidada com o MS2 acima pela diferença de massa entre "picos".

O Orbitrap XL é capaz de dissociar moléculas por três técnicas: dissociação por colisão induzida (do inglês *colision activated dissociation* (CAD)), dissociação por transferência de elétrons (do inglês *Electron Transfer Dissociation* (ETD) (12)), e dissociação por *High Energy Dissociation* (HCD) no C-trap. O CAD é uma das metodologias mais antigas e usadas em proteômica *shotgun*; ela é extremamente eficaz para íons com carga +2 e +3. O ETD, recentemente criado, é indicado para moléculas maiores que adquirem maior carga. Adicionalmente, o ETD é indicado para estudos de moléculas com modificações pós-traducionais por conservar a modificação na molécula durante a sua dissociação (figura 4).

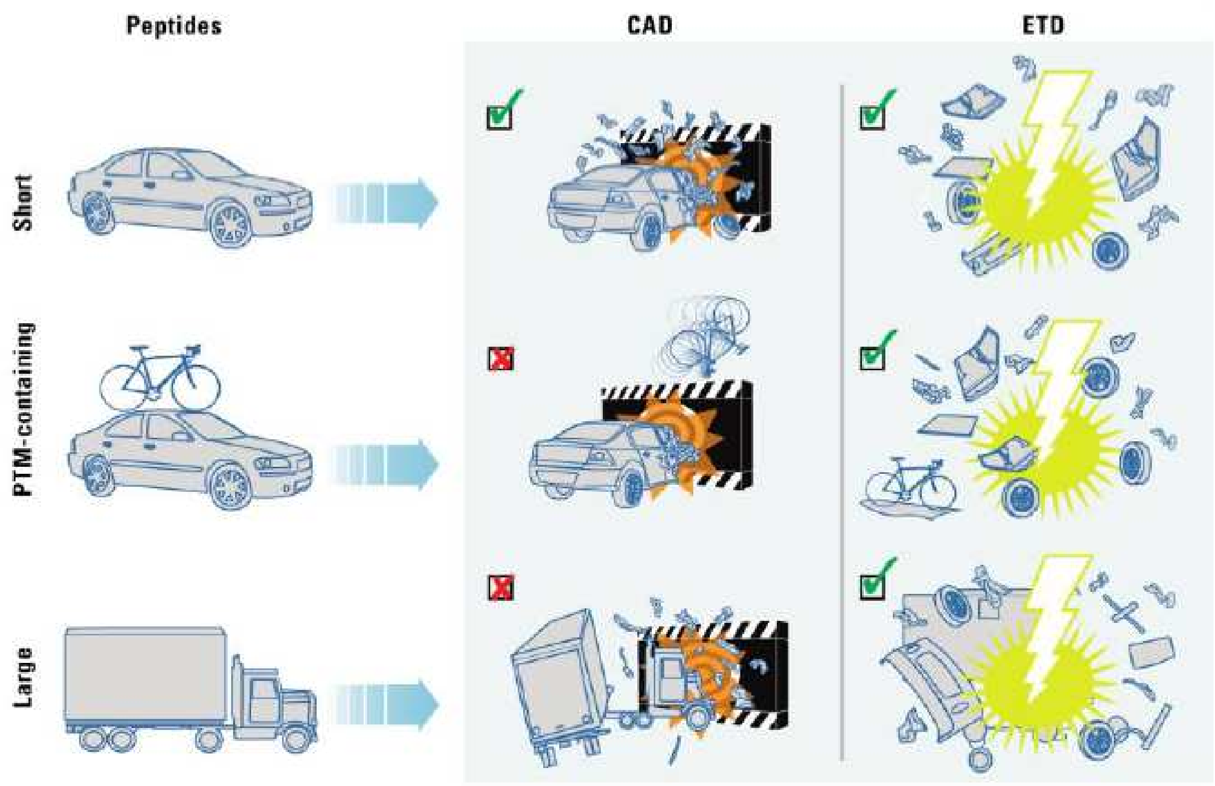


Figura 4 – Dissociação por colisão induzida e por transferência de elétron. Figura obtida de (13). A Figura sugere o ETD como sendo mais eficiente do que o CAD para a análise de moléculas maiores ou com modificações pós-traducionais.

Os analisadores LTQ e Orbitrap podem trabalhar em paralelo onde a configuração mais utilizada é a obtenção do MS1 no Orbitrap, e análises subsequentes de MS2 no *linear trap*. Esta configuração é eficiente, pois toma proveito da alta resolução do Orbitrap para determinar o *m/z* de(s) íon(s) precursor(es) a serem dissociados, e usa a velocidade do LTQ para obter diversos MS2 de íons apontados no MS1. Geralmente, para cada MS1 adquirido no Orbitrap, seis espectros de massa em *tandem* são analisados no LTQ. As seções subsequentes detalham as partes do espectrômetro utilizado.

1.2.2 Ionização por "*spray* de elétrons"

A ionização por *spray* de elétrons, ou *electrospray* (ESI), foi aperfeiçoada para estudos de macromoléculas biológicas por John Fenn et al. em 1989 (14). Esta técnica transfere e ioniza analitos da fase líquida para a gasosa permitindo análise por espectrometria de massa. Métodos de ionização anteriores, tais como, bombardeamento de átomos rápido (FAB) ou dessorção[10] por plasma, provocam a fragmentação do analito e a formação de íons a partir de moléculas neutras, limitando a sua aplicabilidade ao estudo de biomoléculas (15).

Na técnica ESI, solução contendo o analito e eletrólito[11] em baixa concentração (~0,5%) é introduzida na câmara de ionização do espectrômetro via capilar. A injeção é realizada por sistema cromatográfico de micro ou nanofluxo, ou por seringa acoplada a controlador. Uma voltagem de 1 a 5 kV é aplicada entre a extremidade do capilar na câmera de ionização e o orifício de entrada do espectrômetro, tornando-os dipolos de um campo elétrico (figura 5).

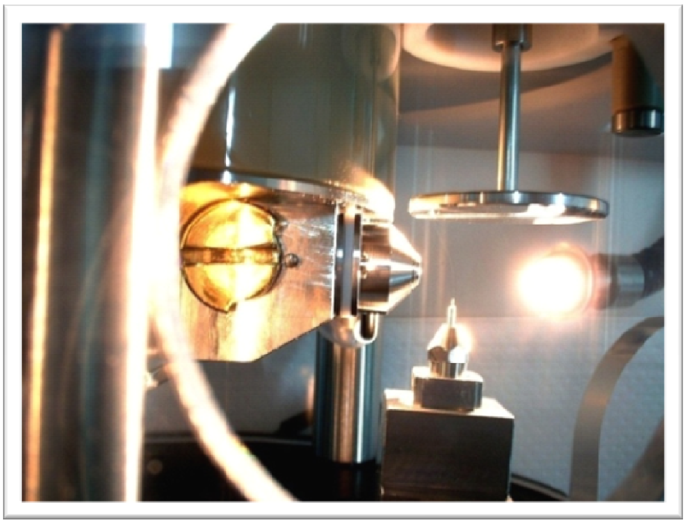


Figura 5 - No espectrômetro Micromass Ultima, o injetor de analito (capilar) é localizado de tal forma que o feixe de íons incida perpendicularmente ao orifício de entrada do espectrômetro.

[10] Fenômeno de retirada de substância(s) adsorvida(s) ou absorvida(s) por outra(s). (Dicionário Houaiss)

[11] Condutor elétrico de natureza líquida ou sólida no qual cargas são transportadas por meio de íons. (Dicionário Houaiss)

Esta diferença de potencial concentra íons na extremidade da gota da solução formada na ponta do capilar por um processo conhecido como eletroforese[12]. O aumento de carga altera a forma esférica da microgota para a forma de um cone, denominado "cone de Taylor". Conforme a densidade de carga na gota aumenta, o campo elétrico entre o capilar e o contra eletrodo se intensifica. Isto desestabiliza o menisco e causa ruptura do cone, transformando-o em um *spray* eletrolítico" de microgotas altamente carregadas que viajam entre a ponta do capilar e o contra eletrodo, sofrendo dessolvatação[13] (figura 6).

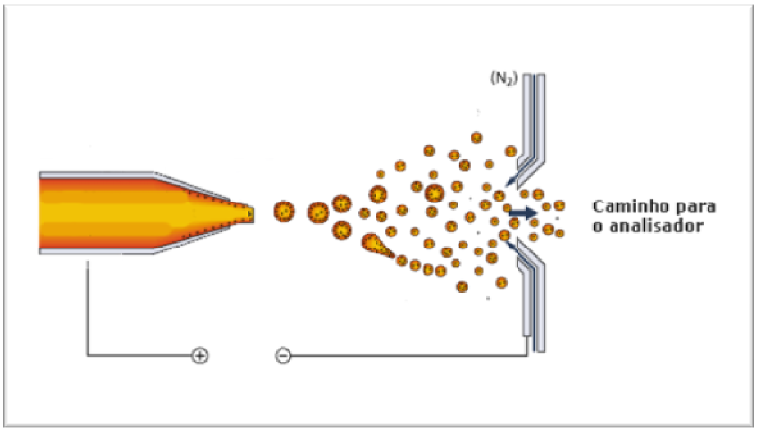


Figura 6 – Microgotas são expelidas do cone de Taylor localizado na ponta do capilar e viajam até a entrada do espectrômetro sofrendo dessolvatação e fissão.

A aplicação de gás secante[14] e energia térmica aceleram a evaporação do solvente, diminuindo o raio da microgota e aumentando as forças eletrostáticas repulsivas que nela atuam. Ao superar o limite de Rayleigh[15], ocorre a "explosão coulombiana", ou fissão da micro-gota "parental", transformando-a em gotículas altamente ionizadas como exemplificado na figura 7(16). Íons analitos são ejetados das microgotas na forma gasosa, por repulsão eletrostática.

[12] Processo de migração para os eletrodos de espécies que são carregados eletricamente em solução (Dicionário Houaiss).

[13] A dessolvatação caracteriza a separação do analito com as moléculas do solvente.

[14] Normalmente nitrogênio é utilizado como gás secante para aumentar a taxa de evaporação do solvente.

[15] O limite de Rayleigh caracteriza o momento que a força de repulsão eletrostática se iguala à força da tensão superficial.

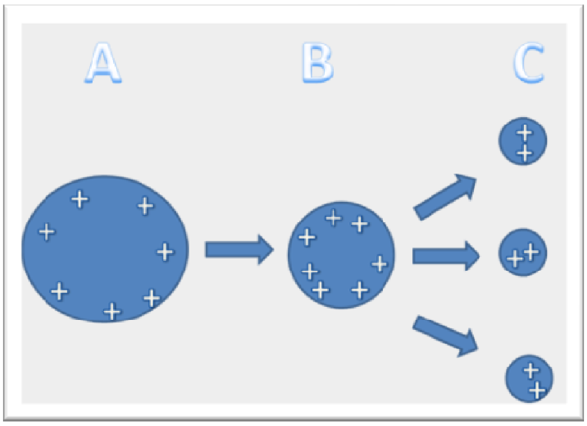


Figura 7 - A) Micro-gotas carregadas são vaporizadas. B) O solvente evapora devido à energia térmica e gás secante, diminuindo o raio da gota e aumentando as forças de repulsão eletrostáticas. C) Quando as forças eletrostáticas superam a força de tensão superficial, ocorre a "explosão coulombiana".

A repetição do processo evaporação e fissão ocorre, inúmeras vezes, simultaneamente durante o trajeto das micro-gotas entre a ponta do capilar e o orifício de entrada do espectrômetro (~ 2 cm), vaporizando e ionizando o analito. A frequência com que o processo se repete é função da intensidade do campo elétrico, da tensão superficial do solvente e da condutividade da solução. A "discretização" deste sinal em pacotes iônicos é necessária para que o analisador separe os íons, e o detector registre o sinal.

1.2.3 O analisador “quadrupolo/armadilha de íons”

Íons oriundos da fonte de ionização são selecionados e guiados até o *ion trap* (do inglês, “armadilha de íons”), por campo elétrico quadrupolar originado de quatro varetas condutoras paralelas, (figura 8). Cada par de varetas opostas é eletricamente conectada para estabelecer o campo eletromagnético; enquanto um dos pares possui potencial $+(U + V.\sin(\omega.t + \emptyset))$, outro possui – $(U + V.\cos(\omega.t + \emptyset))$. Nas equações, U representa um potencial fixo, $V.\cos(\omega.t + \emptyset)$a radiofrequência (RF) com voltagem V, e ωa frequência angular $(2\pi f)$ do potencial de RF dado um ângulo inicial $\emptyset$. A variação senoidal de potencial no tempo entre bastões opostos seleciona íons que possuam trajetória estável (*i.e.* sem a colisão contra os polos) ao longo do filtro. A seletividade do filtro é função da *m/z* do íon, dos potenciais *U*, V, e da frequência ω (figura 8).

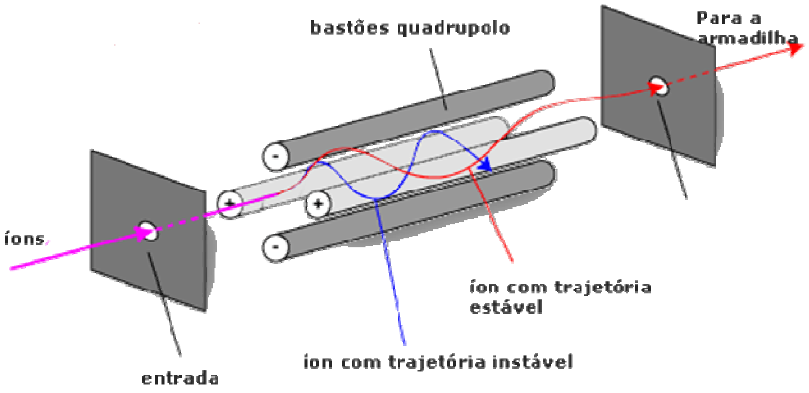


Figura 8 – O quadrupolo filtra e guia íons até a armadilha de íons. O íon representado pelo caminho azul possui trajetória instável e colide com uma das varetas de onde o campo elétrico se origina.

A armadilha de íons utiliza campos eletromagnéticos para aprisionar íons gasosos. Ela é composta por quatro varetas hiperbólicas (figura 9). A dinâmica das partículas em seu interior é descrita conforme a equação diferencial de segunda ordem de Mathieu (17).

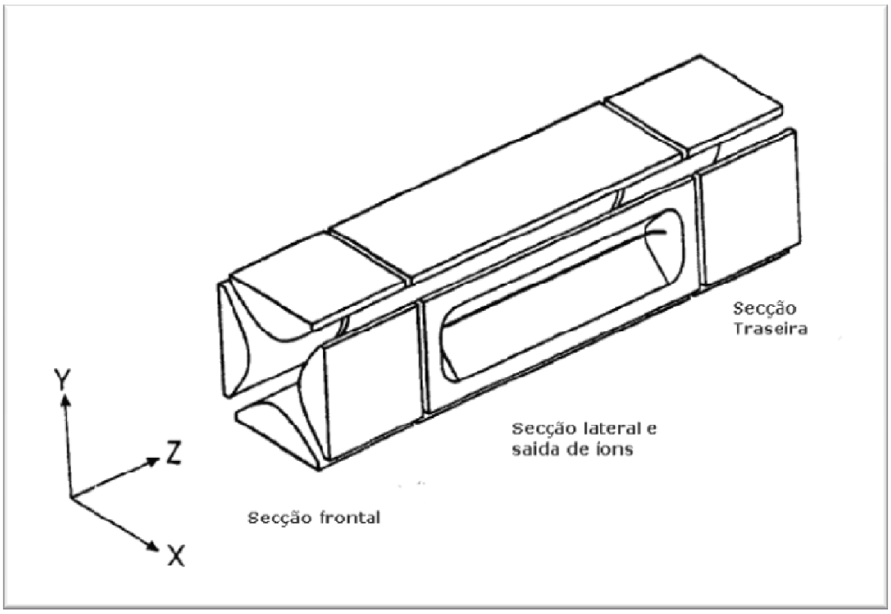


Figura 9 - A armadilha de íons é composta por quatro varetas com lateral hiperbólica.

No espectrômetro LTQ, o MS1 é obtido na armadilha pela separação dos íons de acordo com sua razão massa/carga por excitação ressonante. Resumidamente, íons são aprisionados em órbita ao longo do eixo central da armadilha graças a campos eletromagnéticos de frequência constante (Ex. 1,2 MHz), mas cuja intensidade cresce (Ex. 0 a 10.000 V) com o tempo. O uso de gás hélio para desacelerar os íons que entram na armadilha; os potenciais nas extremidades ajudam no aprisionamento pela rádio frequência. Ao aumentar a intensidade do campo, íons com *m/z* de ordem crescente entram em ressonância com o campo e são expelidos da armadilha para o detector (figura 10).

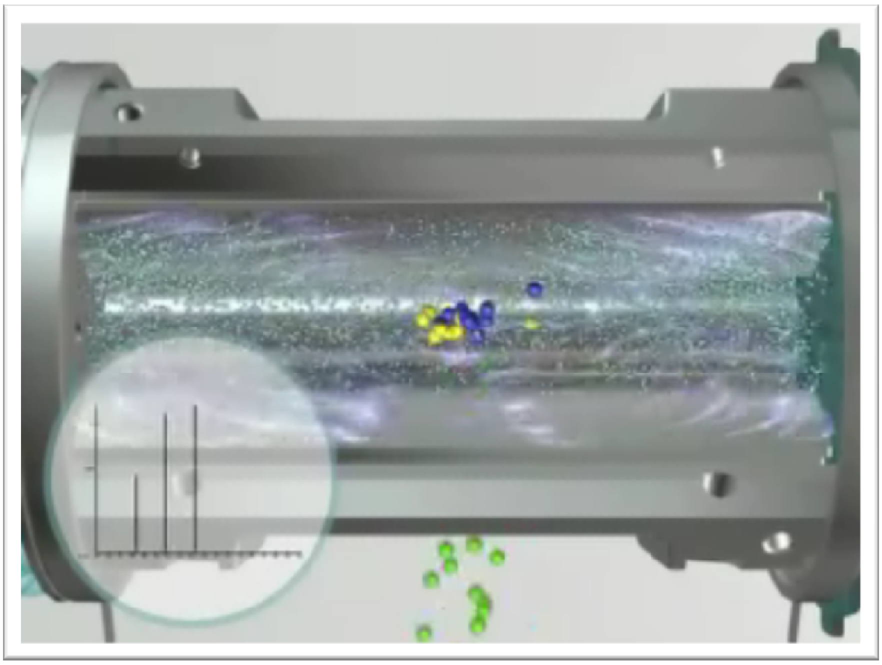


Figura 10 - A ilustração acima é a reprodução de uma animação gentilmente cedida pela Thermo. A figura retrata o interior da armadilha de íons linear. As esferas de diferentes cores representam íons com diferentes m/z e os pontinhos azuis claro, o gás hélio que ajuda o campo elétrico (nuvem azul) a aprisioná-los no centro da armadilha. A imagem retrata o instante em que a frequência/intensidade do campo faz os íons da classe verde entrarem em ressonância e serem expelidos. O lado inferior direito demonstra a geração do espectro (intensidade versus m/z) à medida em que íons são expelidos da armadilha.

O CAD pode ser realizado no *linear trap* aprisionando íons com *m/z* previamente escolhidos e dissociando-os. Resumidamente, é aplicada uma distribuição de frequência/intensidade contendo todas as radiofrequências ressonantes, exceto a dos íons com m/z a serem analisados. Consequentemente, apenas os íons de interesse permanecem na armadilha. Em seguida, a dissociação por colisão induzida é realizada aplicando frequência de ressonância com intensidade insuficiente para ejetar os íons de interesse da armadilha. A energia de colisão entre íons de mesma classe e com o gás hélio é suficiente para dissociá-los. Na próxima etapa, a radiofrequência de ressonância é aumentada progressivamente para expelir o produto de fragmentação dos íons.

1.2.4 O analisador Orbitrap

O analisador Orbitrap (18) pode atingir alta acurácia (uma parte por milhão (ppm) e resolução de 200.000). Um esquema de sua geometria encontra-se na figura 11. Resumidamente, ela é composta de um eletrodo externo (formato de barril) e um coaxial interno (formato de eixo). Juntos, eles geram o campo eletrostático com uma distribuição de potencial “quadro-logarítmica” necessário para garantir uma órbita estável para os íons a serem analisados. Estes, por sua vez, são injetados tangencialmente para dentro do campo produzido, ficando aprisionados, pois a atração eletrostática pelo eletrodo interno iguala-se à força centrípeta. Portanto, os íons seguem uma órbita em torno do eletrodo central. A frequência desta órbita pode ser facilmente demonstrada a ser inversamente proporcional à raiz quadrada da *m/z* dos íons. Como íons de diferentes *m/z* entram no analisador, a frequência orbital destes podem ser determinadas aplicando a transformada de Fourier ao sinal que provém detector.

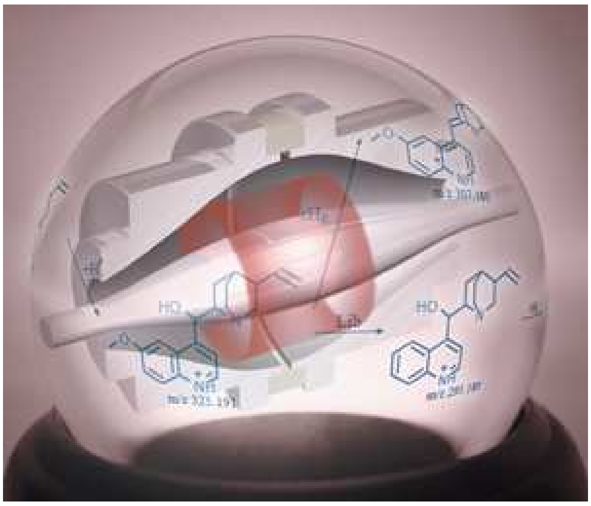


Figura 11 – Esta figura foi obtida no *web site* da Thermo Fisher e representa a geometria do analisador Orbitrap. O anel em vermelho descreve o caminho dos íons durante a análise.

1.2.5 Detector de íons

O detector é considerado o "olho" do espectrômetro de massa por registrar a corrente iônica oriunda do analisador ao longo do tempo. A maioria dos detectores toma proveito do fenômeno de "emissão secundária" para amplificar o sinal de entrada. Este fenômeno é propriedade de apenas alguns materiais condutores e é caracterizado pela emissão de elétrons após a colisão de um elétron com o material emissor. A "multiplicação de elétrons" denomina emissões secundárias sucessivas, e está exemplificada em um tubo capilar de um detector microcanal (MCP) na figura 12. O MCP contém diversos microtubos internamente revestidos com um material capaz de realizar emissões secundárias. Quando o lado de captação do microcanal é atingido por íons, inicia-se o efeito cascata de sucessivas emissões secundárias. Os íons captados e os originados da multiplicação de elétrons são acelerados pela diferença de potencial imposta dentro do microcanal, e a corrente elétrica é registrada.

A análise de misturas complexas de proteínas (Ex. fluidos biológicos ou lisados celulares) gera milhares de íons com diferentes razões *m/z*. A eficiência do detector está associada a sua capacidade de recuperação, ou a taxa de decaimento do potencial, após receber um sinal. Um sinal intenso pode saturar o detector e mascarar o registro de íons menos abundantes. Adicionalmente, em misturas complexas, íons de diferentes classes, mas com a mesma razão *m/z* representam um desafio para serem identificados. Para solucionar estes problemas, a mistura a ser analisada pode ser pré-fracionada por técnicas de cromatografia líquida descritas a seguir.

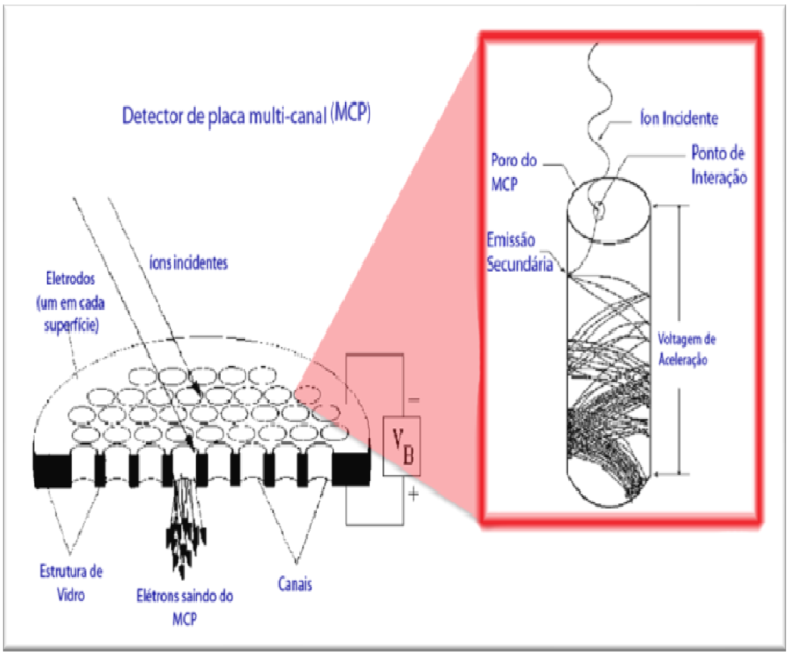


Figura 12 - Ilustração de amplificação de sinal pelo detector *Multi Chanel Plate*. Elétrons incidem no tubo capilar iniciando a multiplicação de elétrons. Estes são acelerados por potencial elétrico de aproximadamente 1.500 V colidindo contra a parede do tubo e liberando novos elétrons (emissão secundária). Este efeito se repete, amplificando o sinal de entrada do microtubo.

1.2.6 Multi-dimensional Protein Identification Technology

A técnica *Multi-dimensional Protein Identification Technology* (MudPIT) tem por objetivo identificar todos os constituintes de uma mistura complexa de peptídeos. O MudPIT utiliza cromatografia líquida de troca iônica (SCX) intercalada com cromatografia de fase reversa (RP) diretamente acoplada a espectrometria de massa em *tandem*.

A cromatografia de troca iônica separa proteínas de acordo com sua carga. Existem diversas matrizes disponíveis comercialmente para este fim; elas são resinas contendo cargas positivas ou negativas que permitem o fracionamento de proteínas de acordo com a distribuição de sua carga superficial. O inicio da separação consiste em aumentar gradualmente a concentração iônica (Ex. concentração salina) durante a lavagem, permitindo eluição diferencial do analito na ordem de menor para maior afinidade com a resina.

A cromatografia de fase reversa (RP) engloba qualquer método cromatográfico que use uma fase estacionária apolar e uma fase móvel polar. Neste método, a fase estacionária é empacotada com hidrocarboneto apolar (ex. C_4, C_8, C_{18}); a fase móvel, por sua vez, contém solvente polares orgânicos (i.e., acetonitrila, butanol, etc.). Quando a mistura de analitos é aplicada, as moléculas mais polares eluirão primeiro, enquanto as menos polares permanecerão ligadas à fase estacionária. A eluição das demais moléculas será alcançada com o aumento da concentração do solvente orgânico, e o tempo de retenção de um componente específico será de acordo com o seu grau de hidrofobicidade (os mais hidrofóbicos serão os últimos).

O MudPIT utiliza nanofluxo no capilar que contem em seu interior resina de SCX seguida de RP. O MudPIT permite separar os componentes da mistura complexa para identificá-los por espectrometria de massa em *tandem* (figura 13). Nesta técnica, a cromatografia de troca iônica decorre aplicando-se durante a eluição uma função degrau de aumento de concentração salina para liberar pacotes de peptídeos da coluna de troca iônica que ficarão retidos na coluna de fase reversa. A cada degrau salino, a coluna RP é submetida ao gradiente hidrofóbico, para eluir os peptídeos ao espectrômetro de massa, e serem identificados por MS/MS. A técnica usa, geralmente, doze degraus de concentração salina, produzindo doze corridas cromatográficas LC/MS de RP (figura 14).

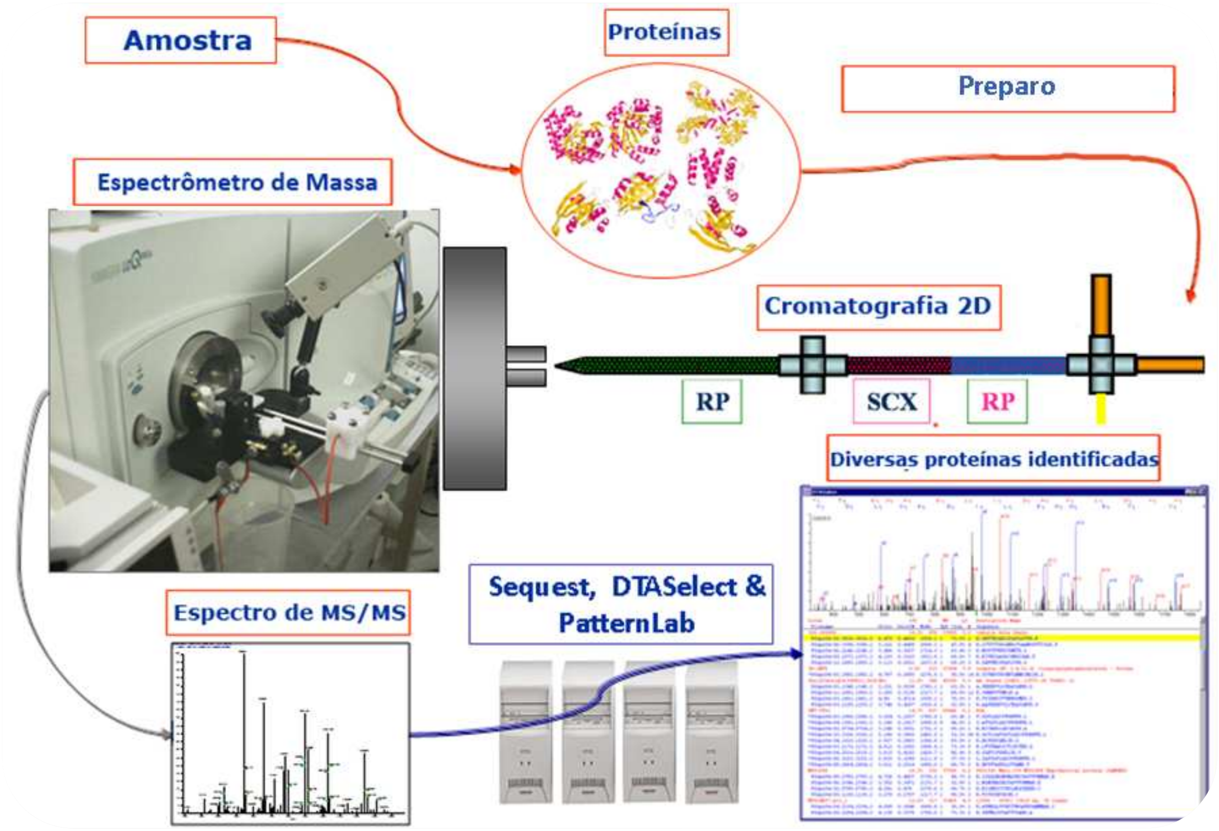


Figura 13 - MudPIT engloba o processo de digestão, separação e identificação dos constituintes de mistura complexa usando cromatografia de nanofluxo com coluna de troca iônica seguida de coluna de fase reversa acoplada diretamente a espectrometria de massa em *tandem*.

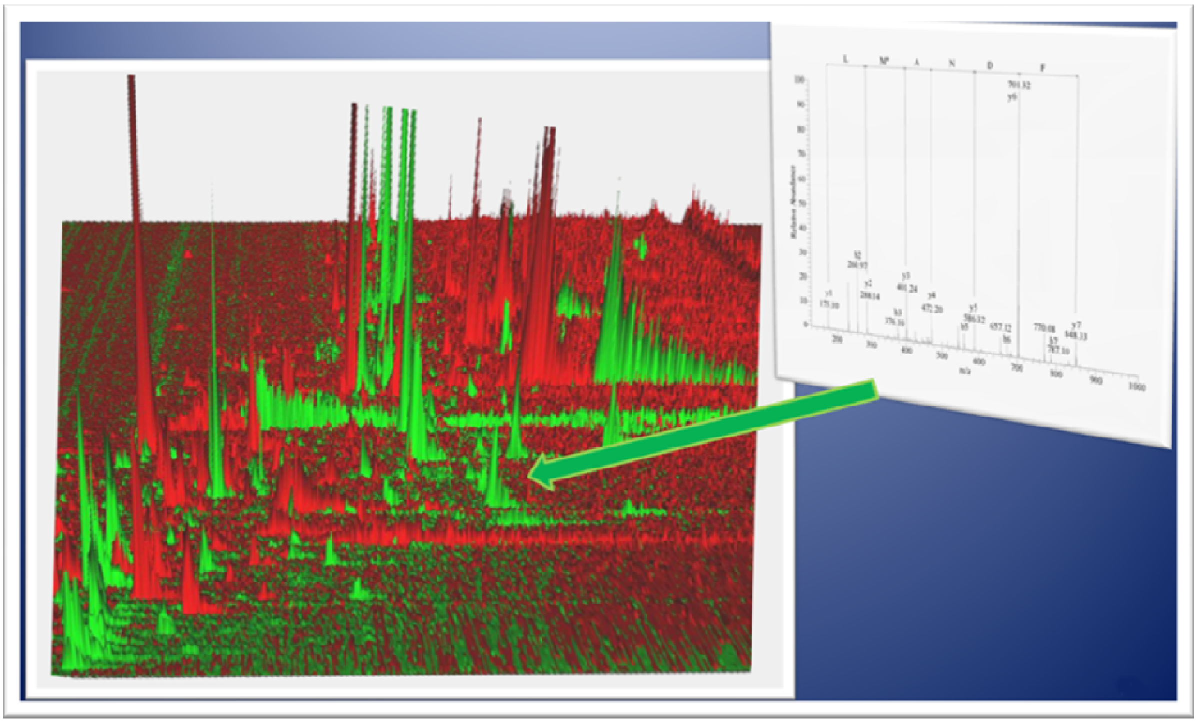


Figura 14– Cada eluato de um “passo salino” aplicado à coluna de troca iônica e seguidamente “separado” por cromatografia liquida de fase reversa (RP). Durante a RP, o eluato é analisado a cada fração de segundo por MS, alternando entre o MS1 e o MS2 de cada um dos quatro íons indicados pelo MS1. O quadro à esquerda representa o “enfileiramento” dos MS1’s obtidos durante um, dos doze passos salinos (x representa m/z, y representa tempo de eluição, e z a intensidade do íon), e o quadro à direita representa o MS/MS de uma das classes de íons indicadas como mais intenso pelo PMF para um determinado instante.

2 Objetivos e problemas abordados

Os subitens abaixo enumeram os objetivos específicos do presente trabalho e contêm uma breve descrição da(s) justificativa(s) que levaram ao desenvolvimento de cada um.

2.1 Objetivo: Apontar biomoléculas diferencialmente expressas entre misturas complexas de biomoléculas

Justificativa: O problema na identificação de proteínas diferencialmente expressas entre misturas complexas é um dos cernes do estudo da proteômica. Do ponto de vista experimental, é desafiador por ser manualmente trabalhoso e exigir equipamentos caros. Na área computacional, implica no desenvolvimento de algoritmos multidisciplinares (espectrometria de massa, bioquímica, reconhecimento de padrões) geralmente "custosos"; nesta tese foi usado cluster com 500 nós contendo processadores Xeon de 2,4 GHz. Adicionalmente, é observada variabilidade na identificação e quantificação proteica devido à natureza da técnica de *shotgun* (19). Devido ao alto custo, as análises geralmente ficam restritas a triplicatas de cada estado biológico estudado; isto torna o problema ainda mais desafiador do ponto de vista de reconhecimento de padrões.

2.2 Objetivo: Auxiliar na interpretação biológica de experimentos de proteômica *shotgun*

Justificativa: Experimentos de proteômica *shotgun* geram listas contendo milhares de proteínas identificadas e quantificadas. A análise dos resultados é extremamente desafiadora, mesmo para especialistas. Portanto, faz-se necessário o desenvolvimento de *softwares* para auxiliar na interpretação dos dados.

2.3 Objetivo: Inferir a carga de íons precursores com base no MS2 obtido por ETD.

Justificativa: Do ponto de vista computacional, o primeiro passo para identificar um peptídeo por espectrometria de massa é restringir o espaço de hipóteses aos candidatos mais prováveis; isso é realizado selecionando-se, de um banco de dados, sequências que correspondem a peptídeos com massa aproximadamente igual massa do íon precursor. A seleção da sequência mais provável é posteriormente realizada por aplicação de métrica de semelhança entre os espectros teóricos dos candidatos e o espectro proveniente da fragmentação do precursor obtido experimentalmente (i.e., espectro de massa em *tandem* ou MS2). Contudo, o espectrômetro de massa fornece somente a *m/z* dos íons, consequentemente, para se conhecer a massa do precursor implica em conhecer-lhe a carga.

A utilização de analisadores de alta resolução (*e.g.*, Orbitrap) possibilita inferir a carga do íon precursor com base no seu envelope isotópico conforme demonstrado na figura 15; a importância da resolução é demonstrada na figura 16. Entretanto, equipamentos de baixa resolução (*e.g.,* LTQ), não possuem poder suficiente para discernir picos de envelopes isotópicos; logo, não permitem estimar a carga de íons peptídeos precursores.

A solução utilizada na identificação proteica, com dados de baixa resolução, em experimentos do tipo *bottom-up*[16], é realizar duas buscas (usando ferramentas como o SEQUEST (20), ProLuCID (21), ou Mascot (22)), uma assumindo que o íon precursor obteve carga +2, e a outra +3. Espera-se que em uma segunda etapa, o(s) resultado(s) falso-positivo(s) serão filtrados através de algoritmo como o DTASelect (23). Existe uma forte correlação direta entre a massa da molécula e a carga

[16] Proteômica *bottom-up* objetiva analisar misturas complexas de proteínas digeridas com tripsina.

adquirida durante sua ionização; quanto maior a massa, maior a carga. A digestão por tripsina garante que o digesto será constituído, em grande maioria, de pequenas moléculas que, assim, adquirirão uma carga baixa (i.e., +2 ou +3); consequentemente, a hipótese de restringir a análise às cargas +2 e +3 é bastante plausível. Vale ressaltar que a metodologia mais usada em proteômica *shotgun* faz uso de analisador LTQ (que possui baixa resolução) e dissociação por CAD.

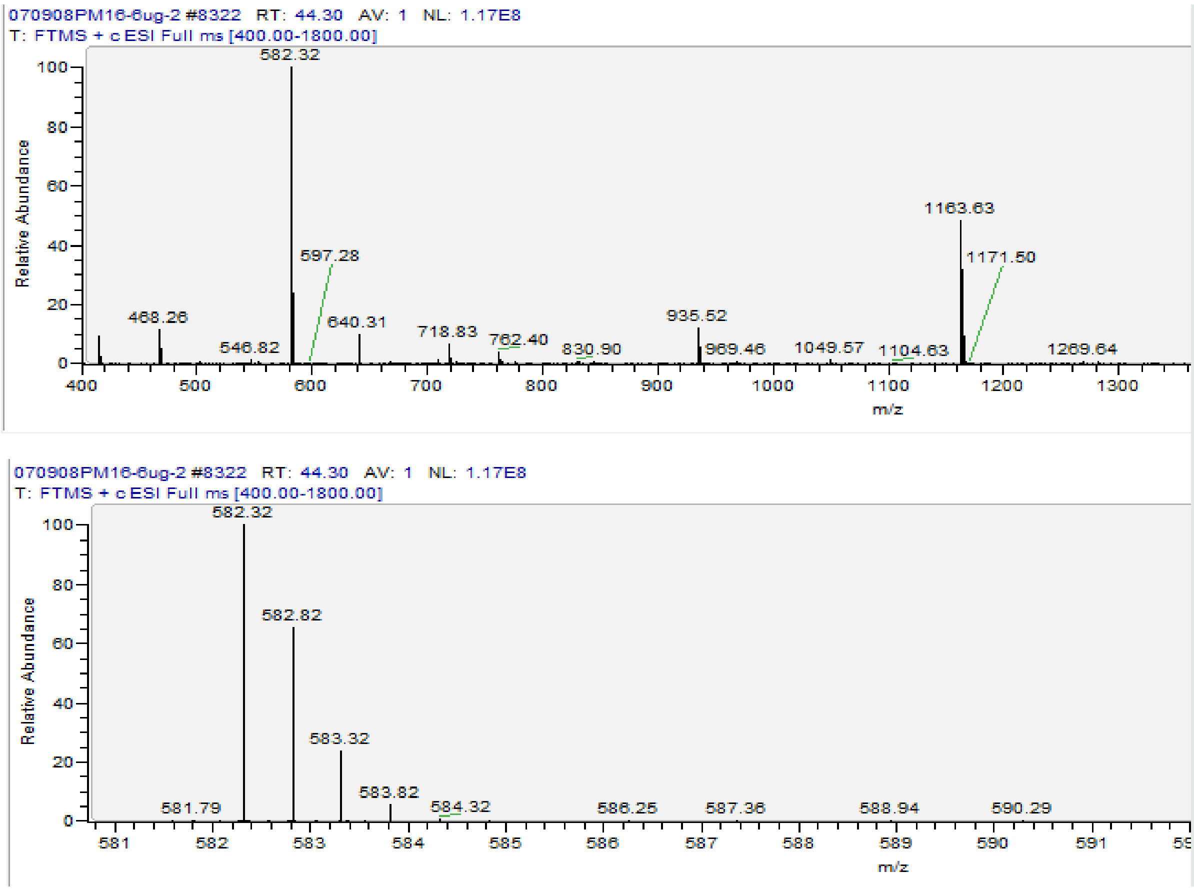


Figura 15- O painel superior e inferior são espectros de massa obtidos com o Orbitrap XL em alta resolução (60.000); o inferior é uma ampliação da região em torno do íon com *m/z* de 582.32. Observa-se um padrão característico de um envelope isotópico no painel inferior onde o íon monoisotópico possui *m/z* de 582.32. É possível inferir a carga do íon pelo espaçamento entre os picos dos isótopos; neste caso de 0.5. Sabendo que a diferença de massa entre cada isótopo deveria ser de aproximadamente 1 Dalton (devido à presença de carbono 13), fica fácil deduzir que este íon adquiriu carga +2 dado que o eixo “X” representa *m/z*.

Experimentos do tipo proteômica *middle-down*[17] objetivam estudar biomoléculas maiores do que as obtidas quando digerindo proteínas com tripsina. A proteômica *middle-down* tornou-se possível, em parte, graças ao recente advento do ETD, que é mais eficiente na dissociação de "moléculas grandes" do que o CAD; adicionalmente, o ETD é capaz de conservar modificações pós-traducionais tornando-se uma ferramenta atraente para estudos, em larga escala, de modificações pós-traducionais (*e.g.*, fosfoproteôma, glicoproteôma, etc.). Em *middle-down*, é comum a observação de íons precursores com carga variando de +2 a +7 ou mais. A geração de maiores possibilidades de estados de cargas dificulta o processo de identificação com dados de baixa resolução. Conforme mencionado acima, cada hipótese de carga geraria uma busca, inflando o número destas em um método custoso e aumentando a inclusão de resultados falso-positivos. Portanto o desenvolvimento de uma estratégia para predizer a carga do íon precursor em estudos de proteômica *middle-down* utilizando espectrômetros de massa de baixa resolução faz-se necessário.

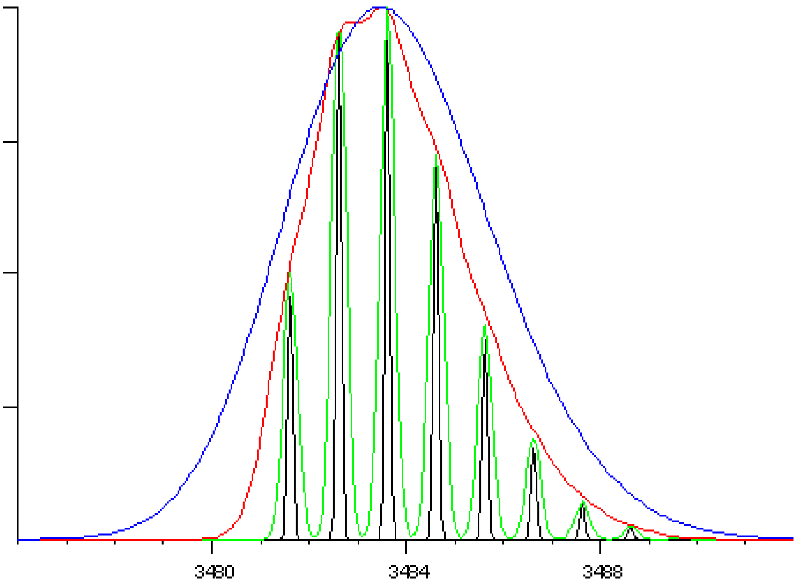


Figura 16 – Esta figura foi obtida do site da MatrixScience. Ela demonstra a visualização de um envelope isotópico com uma resolução de 1.000 (azul), 3.000 (vermelho), 10.000 (verde), e 30.000 (preto). No exemplo acima, fica impossível discernir entre os isótopos quando analisando nas duas menores resoluções.

[17] Entende-se por *middle-down proteomics* experimentos que utilizam enzimas que produzem moléculas maiores do que as geradas pela tripsina.

2.4 Objetivo: Deconvolução[18] e associação da massa monoisotópica dos precursores aos respectivos MS2

Justificativa: A execução da espectrometria de massa em tandem exige o isolamento de íons parentais (precursores) com *m/z* pré-determinado. O algoritmo que rege a seleção de íons precursores no espectrômetro durante a cromatografia usa para a tomada de decisões os picos presentes no MS1 e cujas m/z não se encontram na lista de exclusão dinâmica[19]. Dado um *m/z* de interesse, o espectrômetro é capaz de isolá-los e dissociá-los; geralmente por CAD ou ETD. Embora traga problemas, a tolerância de massa (~1 Da) na seleção do íon parental traz vantagens; dentre elas, a de permitir acumular o sinal dos isótopos do íon selecionado. Isto permitirá o aumento do sinal em relação ao ruído do espectro. Um ponto negativo seria o aumento da probabilidade de seleção de íons de diferentes espécies (oriundos de peptídeos distintos com *m/z* aproximados).

As atuais ferramentas de busca, para identificação proteica, não estão suficientemente desenvolvidas. Consequentemente não permitem desfrutar todas as informações oferecidas pelos espectros de alta resolução (*e.g.*, obtidos com um Orbitrap); um exemplo é a alta resolução que permitirá maior redução no espaço de hipótese das sequências peptídicas candidatas. No MS2, a massa do íon precursor reportada nem sempre é a monoisotópica; contudo as ferramentas de busca consideram apenas a massa monoisotópica das sequência peptídicas no banco de dados. Para contornar esta limitação, as ferramentas de busca são obrigadas a considerar diversas possibilidades de massas monoisotópicas a partir do precursor informado (geralmente 6). Este espaço de busca "inchado" eleva o custo computacional e a possibilidade da inclusão de falso-

[18] Entende-se por deconvolução de espectro de massa o processo de transformar um espectro de massa que contém íons em diversos estados de carga (e.g., +2, +3, ...,+n) em um espectro onde todos os íons contem carga +1.

[19] O processo normal de experimento *shotgun* é uma repetição de MS1 seguido de vários (~6) MS2 durante a cromatografia. O MS1 tem como objetivo prelecionar razões *m/z* a serem submetidos ao MS2. Os íons de maior intensidade são favorecidos pelo algoritmo de escolha. Ao ser submetido ao MS2, o m/z escolhido é incluído em uma lista de exclusão dinâmica onde ali ficará por um tempo predeterminado pelo usuário e não será submetido a um novo MS2 enquanto permanecer nesta lista. Portanto esta técnica tem como objetivo permitir a análise de íons de menor intensidade.

positivos. Adicionalmente, ferramentas de busca não estão preparadas para identificar um espectro MS2 multiplex[20].

2.5 Objetivo: Aumentar o número de peptídeos identificados e a confiabilidade de quantificação nos experimentos *shotgun*.

Justificativa: Aumentar o número de peptídeos identificados e a confiabilidade de quantificação é objeto de constante pesquisa em proteômica.

Apesar das técnicas existentes serem capazes de identificar milhares de constituintes de misturas complexas, os resultados estão aquém de atingir o número verdadeiro de constituintes de misturas complexas, principalmente por dois motivos:

1) **Desafio**: A grande quantidade de constituintes é desafiadora por que a identificação de cada classe de íons requer uma análise de MS2. **Limitação**: Muitas vezes o espectrômetro de massa não é capaz de obter MS2 a tempo de cada uma das diferentes classes de íons que coeluem da corrida cromatográfica.
2) **Desafio**: A ampla diferença na ordem de grandeza (10^12) na quantidade com que os constituintes estão presentes na amostra. **Limitação**: O sinal de íons presentes em maior quantidade mascara o sinal dos íons presentes em menor quantidade.

A quantificação é comumente obtida por integração do cromatograma de íons ou por contagens espectrais. É fácil demonstrar que quanto mais a molécula for identificada em uma análise MudPIT, maior será a confiabilidade da sua quantificação.

Concluindo, devido ao grande número de moléculas coeluindo durante uma análise MudPIT, o pesquisador deve assumir um compromisso entre número de identificações únicas e a confiabilidade na quantificação.

[20] O espectro multiplex é gerado pela co-fragmentação de dois ou mais íons de diferentes espécies.

3. Metodologia e resultados

Os resultados desta tese encontram-se apresentados na forma de artigos publicados. As seções abaixo têm a finalidade de introduzir os artigos que abordam os objetivos específicos propostos, respectivamente. No caso dos artigos de livre acesso, estes se encontram reproduzidos na integra. Para os demais, foi reproduzido apenas o resumo.

3.1 Apontar biomoléculas diferencialmente expressas entre misturas complexas de biomoléculas

Os resultados de experimentos MudPIT geram listas contendo milhares de proteínas identificadas e quantificadas. Os primeiros experimentos desta tese voltaram-se à avaliação de métodos estatísticos para apontar proteínas diferencialmente expressas entre misturas complexas analisadas por MudPIT. Para isso, foi criado um experimento real, mas controlado, onde foram adicionadas proteínas marcadoras em solução com lisado de levedura. Os marcadores foram adicionados em diferentes concentrações em diversos lisados, criando assim conjuntos de validação ou gabaritos onde sabe-se, de antemão, quais são as proteínas diferencialmente expressas. Estes gabaritos foram usados para avaliar se metodologias de reconhecimento de padrões seriam capazes de apontar quais os marcadores quando comparados dois lisados onde os marcadores encontravam-se em diferentes concentrações. Os marcadores utilizados foram: soro albumina (ALB), fosforilaze b (PHS2), inibidor de tripsina (ITRA) e anidrase carbônica (CAH) da Bio-Rad. Estes foram adicionados a quatro alíquotas iguais de lisados de levedura (400 µg cada) em concentrações relativas de 25%, 2,5%, 1,25% e 0,25% do lisado. As amostras foram digeridas com endoproteinase Lys-C e tripsina.

No primeiro conjunto de validação, rotulamos como classe +1 as três leituras (de MudPIT) das amostras contendo 25% de proteínas marcadoras em relação à massa proteica do lisado de levedura. As outras leituras, referentes às concentrações 2,5%, 1,25% e 0,25%, foram rotuladas –1 (+25, | –2,5, –1,25, –0,25). O segundo conjunto rotula os dados das concentrações de 25% e 2,5% como +1 e os restantes como –1 (+25, +2,5 | –1,25, –0,25). Finalmente, o terceiro conjunto rotula apenas os dados da concentração de 0,25% como pertencentes à classe negativa (+25, +2,5, +1,25 |–0,25). As metodologias computacionais avaliadas para apontar as proteínas diferencialmente expressas e os resultados deste experimento estão descritas no **Artigo I ("Identifying differences in protein expression levels by spectral counting and feature selection")** do anexo.

Os algoritmos criados para a realização do Artigo I são de fácil aplicabilidade e grande utilidade para uso em laboratórios de proteômica. Motivado pela aplicabilidade, criou-se ambiente com interface gráfica que reunia os *parsers*, metodologias de normalização, e de reconhecimento de padrões utilizadas. Este ambiente foi denominado de *PatternLab for proteomics*.

O *PatternLab* logo ganhou dois módulos com novas funções de reconhecimento de padrões. O primeiro módulo, intitulado nSVM (de *natural support vector machines*) combina algoritmo genético (AG) com máquinas de vetor de suporte (SVM). O seu objetivo é apontar um conjunto mínimo de proteínas que estão diferencialmente expressas entre misturas complexas de tal forma a otimizar uma função classificadora. Este conjunto mínimo seria as proteínas mais indicadas para a construção de kits de diagnósticos.

Algoritmos genéticos (AG) são métodos computacionais, inspirados na natureza, que buscam o ponto ótimo para uma função matemática (função objetivo) simulando processos naturais de evolução. O AG inicia-se criando população aleatória de "cromossomos" (estruturas de dados representando possíveis soluções ao problema) que "evoluem" até satisfazer o critério de convergência pré-estabelecido. Aqui se compreende por "evolução" a seleção de genitores com probabilidade proporcional à sua aptidão (definido pela função objetivo), que gerarão sucessor(es) através de cruzamento. Os sucessores sofrerão mutações, em seu genoma, que poderão ser transmitidas aos seus descendentes. Espera-se que, a cada nova geração, os indivíduos tornem-se mais aptos a fornecer melhores soluções à função objetivo. No AG clássico, ao atingir o critério de convergência, a solução encontra-se codificada no cromossomo do indivíduo mais apto.

No nSVM, cada "cromossomo" (indivíduo) constitui um vetor de bits[21] e representa uma solução no espaço de hipóteses[22] de um conjunto de características (proteínas) diferencialmente expressas entre os grupos de amostras (classes). A ligação entre o AG e a problemática da busca dos marcadores é realizada pois o bit (*lócus*) de argumento *n* de um cromossomo representa a

[21] O bit pode assumir apenas os valores zero ou um.

[22] O espaço de hipóteses representa todas as possíveis combinações de proteínas detectadas como sendo as diferencialmente expressas. Caso 1.000 proteínas forem detectadas, o número de combinações existentes é 2^{1000}.

expressão diferencial (bit = 1) ou não (bit = 0) da proteína de índice *n* em sua solução. É observado que o número de *loci* (bits) em cada indivíduo se iguala ao total de proteínas identificadas por MudPIT. Na solução ótima (ótimo global), apenas os bits correspondentes aos índices das proteínas diferencialmente expressas assumem o valor unitário. Entende-se por ótimos locais as soluções que atingiram o critério de convergência, que contudo não representam corretamente as proteínas diferencialmente expressas na solução a ser analisada. Durante a geração de nova população, os sucessores herdarão cada *lócus* com probabilidade de 50%. A introdução de nova(s) mutação(s) tem como objetivo garantir maior varredura do espaço de hipóteses e evitar convergência prematura do AG para ótimos locais. A mutação é efetuada através da alternância aleatória do(s) valor(s) de bit(s) nos novos indivíduos.

A aptidão é determinada pelo cômputo da função objetivo. O valor numérico da aptidão de cada individuo do AG é calculada a partir dos dados normalizados de quantificação relativa das proteínas apontadas pelo cromossomo artificial correspondente (*lócus* / bits de valor 1). Estes dados mapeiam os vetores de entrada de um conjunto de validação para um espaço de características; neste, são avaliados: o risco empírico[23] (medido pelo método de validação cruzada *leave-one-out* (*LOO*)), a dimensão Vapnik-Chervonenkis[24] (VC ou *h*), número de vetores de suporte (*nSV*)) necessários para definir o hiperplano separador de máxima margem entre classes e a dimensionalidade da solução (*nAl*). A função aqui usada para cálculo da aptidão (*A*) foi

$$A = C_1 * LOO + C_2 * \left(1 - \frac{1}{h}\right) + \ C_3 * \left(1 - \frac{1}{nSV}\right) + \ C_4 * nAl$$

onde C_1, C_2, C_3 C_4 são constantes. Como quatro proteínas foram adicionadas ao lisado, espera-se que a solução ótima se encontre em espaço de características com quatro dimensões e, estas últimas, correspondam aos índices dos marcadores adicionados. A estimativa de quantos marcadores estavam na amostra era avaliada através de uma lista de duas colunas onde a primeira

[23] Algumas vezes referido na literatura como erro de treinamento, o risco empírico mede o erro intrínseco do conjunto de dados. Este é geralmente calculado empregando metodologia de validação cruzada.

[24] Resumidamente, a dimensão VC é uma medida de complexidade da disposição dos dados no espaço de características, diretamente relacionada com a capacidade de generalização teórica de uma função classificadora. No caso dos classificadores SVM, a estimativa da dimensão VC é função do raio da menor hiperesfera que circunscrever os vetores de suporte, da margem separadora entre classes, tamanho amostral e um fator de certeza.

representa o índice das proteínas e a segunda a frequência no genoma dos mais aptos. Esta lista era ordenada de forma decrescente de frequência (segunda coluna). O número de marcadores foi estimado em dois passos:

1. Localização das duas fileiras consecutivas que apresentam maior diferença entre valores das frequências dos marcadores.
2. A estimativa do número de marcadores será igual ao número de linhas acima do ponto estabelecido no item anterior.

Quando o nSVM foi aplicado ao gabarito gerado para o estudo anterior, ele foi capaz de apontar quais e quantos foram os marcadores adicionados no lisado de levedura.

O segundo módulo adicionado ao *PatternLab*, intitulado ACFold, fornece um "diagrama de expressão proteica global". Este mapeia proteínas a um gráfico de acordo com a alteração na quantificação da expressão, o valor do teste estatístico AC (24), e valor do estimador teórico de BH para falso-positivos(25). O ACFold pode ser empregado para comparar análises MudPIT (tem como vantagem não requerer análises replicatas de cada classe).

Uma descrição do *PatternLab* dando ênfase a estes dois novos módulos encontra-se descrito nos **Artigo II ("PatternLab for proteomics: a tool for differential shotgun proteomics")** e **Artigo III ("PatternLab for differential shotgun proteomics")** do anexo.

Com o passar do tempo, o *PatternLab* começou a ser adotado por grupos externos ao nosso, e com isso, surgiram sugestões de novos módulos. Dentre as sugestões, as mais pedidas foram para criar um módulo capaz de gerar diagramas Venn com área proporcional e um módulo para análise de "experimentos temporais". Para atender aos pedidos, criamos três novos módulos; o "*Approximate area proportional Venn diagram*" (AAPVD), o TrendQuest, e o XFold. O TrendQuest tem como objetivo agrupar proteínas que obtiveram padrão de expressão similar durante um experimento temporal. O XFold estende o conceito do ACFold e do TFold[25]; ele agrupa proteínas que se encontram diferencialmente expressas em intervalos de tempo comuns; por exemplo, proteínas que foram consideradas diferencialmente expressas nos instantes 1, 5, e 13 de um experimento temporal seriam agrupadas. Demonstramos a utilização destes novos módulos em

[25] O TFold segue os mesmos princípios do ACFold mas utiliza o teste t ao invés do AC; ele é recomendado para quando se tem leituras replicatas de cada classe

estudo, onde analisamos os efeitos de um quimioterápico, o álcool perílico, em cultura de células de glioblastoma multiforme. O experimento, resultados e descrição dos novos módulos encontram-se descrito no **Artigo IV (“Dynamic overview of glioblastoma cells exposed to perillyl alcohol”)** do anexo. O **Artigo V (“Analyzing shotgun proteomic data with PatternLab for proteomics”)** do anexo serve como “manual de referência” do PatternLab, além de discutir sugestões de variáveis a serem utilizadas nos diversos métodos e fornecer explicações de como interpretar os resultados. O **Artigo VI (“PYR/PYL/RCAR family members are major in-vivo ABI1 protein phosphatase 2C-interacting proteins in Arabidopsis”)** do anexo tem enfoque puramente biológico; ele encontra-se citado nesta seção como exemplo do uso do *PatternLab* para apontar proteínas de interesse (i.e., diferencialmente expressas). Este artigo foi realizado em colaboração durante a elaboração desta tese.

3.2 Auxiliar na interpretação biológica de experimentos de proteômica *shotgun.*

Experimentos de proteômica *shotgun* geram listas contendo milhares de proteínas identificadas e quantificadas de difícil interpretação. Para auxiliar especialistas na análise destes resultados, criamos um módulo para o *PatternLab* intitulado "*Gene Ontology Explorer*" (GOEx). O GOEx utiliza os dados de anotação do *Gene Ontology* (26) para agrupar proteínas por processos biológicos, componentes celulares, ou funções moleculares. O algoritmo é capaz de mapear dados experimentais ao GO e localizar termos do GO que se encontram estatisticamente sobre-representados.

O *Gene Ontology Consortium* (GO) é um grande projeto que envolve diversos centros acadêmicos com o objetivo de fornecer um conjunto de vocabulários estruturados (termos) usados para descrever produtos de genes em um organismo. Existem três subontologias no GO associadas a produtos de genes; são eles: processos biológicos, componentes celulares, e funções moleculares. Cada subontologia e composto de termos onde cada um pode estar associado a um ou mais termos seguindo a estrutura de um grafo acíclico direcionado. Uma importância prática do GO é conter milhares de termos com anotações para dezenas de organismos.

O GOEx é o primeiro algoritmo a utilizar o GO para auxiliar na interpretação biológica de dados quantificados por proteômica *shotgun*. Uma descrição do algoritmo e exemplo de sua utilização para auxiliar na interpretação dos efeitos do álcool perílico em células de glioblastoma encontram-se descritos no **Artigo VII ("GO Explorer: A gene-ontology tool to aid in the interpretation of shotgun proteomic data")** do anexo.

3.3 Inferir a carga de íons precursores com base no MS2 obtido por ETD.

A inferência da carga do íon precursor é necessária para a identificação proteica. Existe relação direta entre o tamanho da molécula e a carga adquirida no processo de ionização. Os experimentos de proteômica *middle-down* objetivam estudar moléculas maiores do que as geradas após tripsinização. Por isso são utilizados eletivamente no estudo de modificações pós-traducionais; contudo, por envolver moléculas maiores, os estados de carga também são maiores do que quando comparados aos estudos de proteômica *bottom-up*.

Em espectros de baixa resolução, a carga de íons precursores não pode ser inferida, no MS1, pelo envelope isotópico (ex., obtidos com o LTQ); consequentemente, o desafio para a identificação do espectro, especialmente em dados de *middle-down*, é considerável. Para solucionar este problema, aqui é apresentada nova abordagem computacional capaz de inferir a carga do precursor a partir do espectro de massa em *tandem* de baixa resolução (~1000) obtido por ETD. O método, denominado *Charge Prediction Machine* (CPM), utiliza a técnica de "aprendizagem supervisionada" que permite ao algoritmo realizar "aprendizagem estatística dos padrões" utilizando exemplos de um gabarito. O CPM aprende a inferir a carga com base em três características do MS2: os precursores de carga reduzida[26], os íons complementares[27] e as perdas neutras[28]. O gabarito para aprendizagem supervisionada foi criado a partir de lisado celular, com

[26] Em MS2 obtidos por ETD, é comum observar precursores de carga reduzida. Por exemplo, supomos que o íon precursor obteve carga +4 e m/z de 1000. No MS2, esperaríamos encontrar um pico com m/z igual como se o precursor tivesse obtido carga +3 (~1333 m/z) , outro com +2 (~1999 m/z) , e outro como +1 (~3997 m/z). Provavelmente, devido à configuração do espectrômetro de massa, o precursor de carga reduzida de +1 não seria observado no espectro de massa pois o intervalo m/z analisado é geralmente entre 500 m/z ate 2000 m/z.

[27] Entende-se por íons complementares os pares de íons de um espectro em *tandem* cuja soma das massas iguala-se a do íon precursor.

[28] São "perdas neutras", quando uma molécula perde um fragmento que não contém carga. Geralmente esta perda é uma molécula de água, amônia, ou alguma outra molécula de baixo peso molecular. Na problemática desta tese, uma ao observar um pico de precursor de carga reduzida de +2, esperaríamos ver um pico com aproximadamente 9 m/z's a menos que corresponderia a uma perda neutra de H_20 (massa molecular de ~18 Da). A diferença de 9 Da entre o pico da perda neutra e o de precursor de carga reduzida é sugestiva de que o precursor de carga reduzida, de fato, contém carga +2.

uso do Orbitrap para obtenção dos MS1 (de alta resolução) e do LTQ (com dissociação por ETD) para adquirir os MS2 (de baixa resolução). Os dados do Orbitrap garantem a determinação da carga do íon precursor.

O CPM possui em seu cerne um classificador baseado em função discriminante bayesiana, com um "parâmetro de relaxação", por nós introduzido, visando atender o problema adjacente. Para demonstrar este parâmetro de relaxação, o seguinte exemplo é fornecido: nele, o espaço de hipótese de cargas é de +2 a +7 e existem 1.000 espectros em *tandem* e a carga dos precursores precisam ser estimada. A resolução do problema implicaria 6.000 buscas, caso todos os estados de carga fossem considerados pela ferramenta de busca; o número "inflado de buscas" demandará grande poder computacional e aumentará a chance de falso-positivos.

Visando reduzir o espaço de busca e solucionar este problema, os dados são processados através do CPM antes do início das buscas; isto atribuirá aos espectros apenas a(s) carga(s) mais prováveis. Nesta etapa, o CPM utiliza o classificador bayesiano para atribuir uma pontuação a cada estado de carga de cada espectro; quanto menor a pontuação, mais provável a veracidade do estado de carga. Em seguida, estas pontuações serão recalculadas em cada espectro onde a de menor valor assumirá o valor zero, a de segundo menor valor, assumirá a diferença entre seu valor e o do original de menor valor, e assim sucessivamente. Cada espectro terá, consequentemente, um estado de carga onde o de pontuação zero será o mais provável. Na segunda etapa, o algoritmo irá agrupar na memória do computador 6.000 estruturas de dados, doravante referidas como objetos, onde cada um possuirá informação de seu estado e pontuação de carga, com a respectiva referência espectral. Esta lista de objetos é ordenada por pontuação de forma não decrescente implicando que os primeiros 1.000 objetos possuam referências espectrais distintas. A pontuação da menor diferença global entre os dois estados de carga de menor valor, pertencentes ao mesmo espectro, é atingida no objeto de número 1001. A partir daí, é esperado que este objeto referencie o espectro onde o CPM teve maior "dificuldade" de atribuir a carga e inclui a segunda hipótese de carga mais provável para este espectro.

Caso o usuário determine valor 1 ao parâmetro de relaxamento, o CPM usará os primeiros 1.000 objetos da lista para atribuir estado de carga aos espectros; cada espectro receberá, implicitamente, uma hipótese de estado de carga. Se o usuário atribuir o parâmetro de relaxação igual a dois, o CPM atribuirá 2.000 hipóteses de carga aos 1.000 espectros; cada espectro receberia ao menos uma hipótese, contudo neste exemplo, devido à natureza do algoritmo, este número poderá variar até seis. Isto introduz visão holística de otimização ao processo de atribuição de hipóteses de carga. Como é esperado que hipóteses erradas serão filtradas em etapa posterior (por abordagem ortogonal como o DTASelect) será válido admitir que o CPM atribua múltiplas hipóteses de carga a cada espectro.

Nossos resultados mostram que a redução no espaço de hipóteses gerada pelo CPM diminui o número de identificações falso-positivas levando à diminuição drástica do trabalho computacional para identificação proteica. Foi mostrado também que a introdução do parâmetro de relaxação na abordagem permitiu identificar espectros multiplex.

Além dos atributos descritos acima, a comparação do CPM com o único método existente (Charger, Thermo Fisher, San José – CA), mostrou acerto de 98% do primeiro em um gabarito gentilmente cedido pelo laboratório do Dr. Joshua Coon, enquanto o Charger alcançou apenas 86%. A descrição minuciosa do algoritmo do CPM e dos resultados obtidos encontra-se no **Artigo VIII (“Charge Prediction Machine: A Tool for Inferring Precursor Charge States of Electron Transfer Dissociation Tandem Mass Spectra”)** do anexo.

3.4 Deconvolução e associação da massa monoisotópica dos precursores aos respectivos MS2.

Algoritmos para deconvolução de espectros de alta resolução constituem a base para criação de ferramentas especializadas em espectros de alta resolução; os poucos disponíveis não disponibilizam seu código fonte. Nesta seção é introduzida uma ferramenta, por nós desenvolvida, intitulada YADA. Esta ferramenta possui um algoritmo de deconvolução (DA), importante para reprocessar dados do MS2, garantindo que somente a massa monoisotópica do(s) seu(s) íon(s) precursor(es) seja(m) informada(s). Esta inovação permitirá a redução do espaço de busca das ferramentas de identificação proteica. A consequência direta será encurtamento do tempo de identificação espectral de proteínas e diminuição de resultados falso-positivos. A redução do espaço de busca é especialmente crítica em estudos de proteômica *middle-down* e *top-down* por envolverem “moléculas maiores” e, consequentemente, envelopes isotópicos com maior número de picos.

Resumidamente, o algoritmo YADA deconvolui espectros seguindo as etapas abaixo:

1) Os picos espectrais oriundos de “ruído químico” são eliminados usando a estratégia demonstrada na figura 17.

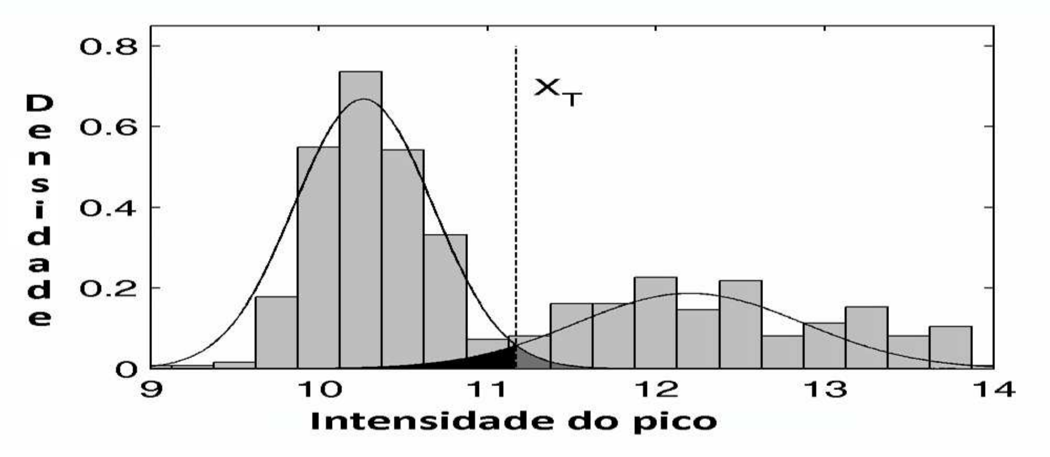


Figura 17 – A intensidade dos picos do MS1 é modelada conforme mistura de duas funções de densidades de probabilidades Gaussianas por algoritmo de *Expectation Maximization*. Em seguida, o classificador ótimo Bayesiano é calculado (X_T). Picos com intensidade inferior a X_T serão classificados como oriundos de ruído químico e serão eliminados.

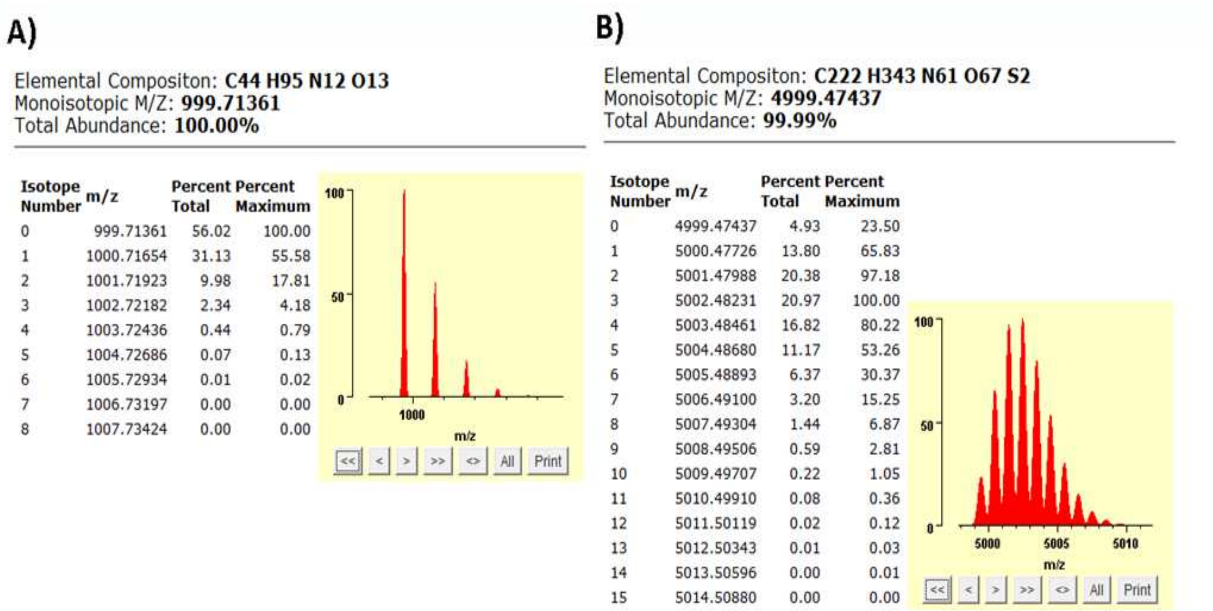
2) A lista de envelopes isotópicos candidatos é obtida. Para isso, cada candidato é caracterizado por grupos de picos que se distanciam, seguindo um padrão de distâncias; exemplificando, para um íon de carga 1 as distâncias serão 1/1,007, de carga 2, 1/(2* 1,007), e assim sucessivamente.

3) Envelopes isotópicos candidatos que não apresentarem perfil de intensidades condizentes serão eliminados. Estes são caracterizados por exibirem menos de 3 picos ou $\langle \boldsymbol{a}, \boldsymbol{b} \rangle < c$ onde $\|\boldsymbol{a}\| = 1, \|\boldsymbol{b}\| = 1$; $\boldsymbol{a}$**,** representa distribuições de intensidade; $\boldsymbol{b}$**,** a intensidade do modelo Averagine predita para a massa em questão e c, uma variável estabelecida pelo usuário. Vale ressaltar que a distribuição da intensidade dos picos do envelope é função da massa do "íon peptídeo" e, que esta distribuição pode ser teoricamente calculada de acordo com o algoritmo Averagine (27). A figura 18 exemplifica como o número e intensidades dos picos de um envelope isotópico variam em relação à massa molecular.

4) Para cada envelope isotópico da lista, é inferida uma carga respectiva de seus íons.

5) Os envelopes da lista são recalculados para que todos os íons exibam carga +1.

Finalmente, o espectro deconvoluido é elaborado; neste, os envelopes da lista serão representados apenas pelo *m/z* de seu monoisotópico e com intensidade igual à integral de seu sinal.



A)

Elemental Compositon: **C44 H95 N12 O13**
Monoisotopic M/Z: **999.71361**
Total Abundance: **100.00%**

Isotope Number	m/z	Percent Total	Percent Maximum
0	999.71361	56.02	100.00
1	1000.71654	31.13	55.58
2	1001.71923	9.98	17.81
3	1002.72182	2.34	4.18
4	1003.72436	0.44	0.79
5	1004.72686	0.07	0.13
6	1005.72934	0.01	0.02
7	1006.73197	0.00	0.00
8	1007.73424	0.00	0.00

B)

Elemental Compositon: **C222 H343 N61 O67 S2**
Monoisotopic M/Z: **4999.47437**
Total Abundance: **99.99%**

Isotope Number	m/z	Percent Total	Percent Maximum
0	4999.47437	4.93	23.50
1	5000.47726	13.80	65.83
2	5001.47988	20.38	97.18
3	5002.48231	20.97	100.00
4	5003.48461	16.82	80.22
5	5004.48680	11.17	53.26
6	5005.48893	6.37	30.37
7	5006.49100	3.20	15.25
8	5007.49304	1.44	6.87
9	5008.49506	0.59	2.81
10	5009.49707	0.22	1.05
11	5010.49910	0.08	0.36
12	5011.50119	0.02	0.12
13	5012.50343	0.01	0.03
14	5013.50596	0.00	0.01
15	5014.50880	0.00	0.00

Figura 18 – Painel A mostra envelope isotópico teórico com resolução de 10.000, do "íon peptídeo" de massa 1.000 Daltons com carga +1 obtido segundo o algoritmo Averagine. O Painel B mostra envelope isotópico teórico de "peptídeo íon" com massa 5000. Vale observar que no envelope de menor massa, o pico mais intenso é o do monoisótopo diferentemente do de maior massa. Observa-se também maior número de picos no envelope de maior massa. Para massas maiores ainda, o pico monoisotópico pode ficar abaixo da sensibilidade de detecção do espectrômetro de massa. Para massas maiores ou íons com cargas mais elevadas, faz-se necessária maior resolução para discernir entre os picos provenientes dos isótopos.

Como o cômputo de um modelo Averagine é custoso, para acelerar nosso algoritmo, foi criado vetor de 25 interpoladores do tipo *Akima Spline*; este último permite modelagem eficiente de mudanças abruptas dos dados (*e.g.*, uma função que transita de um valor constante para uma função logarítmica). Para isso, primeiramente, foram calculados modelos Averagine para massas variando de 500 a 35.000 Da com intervalos de 500. As intensidades do primeiro pico do envelope teórico foram utilizadas na construção do regressor que predirá a intensidade do pico 1; as do segundo pico, para a do regressor que predirá as do pico 2, e assim sucessivamente. Assim, será possível predizer a intensidade dos 25 primeiros picos do envelope isotópico para uma dada massa. Nossos resultados evidenciaram que o vetor de interpoladores é capaz de predizer distribuições isotópicas aproximadamente 1000% mais rápido do que o algoritmo Averagine. Em experimentos do tipo *middle-down* as moléculas presentes não são suficientemente grandes para justificar a presença de vetor contendo mais de 25 regressores.

O YADA também é capaz de reconhecer sobreposição de envelopes isotópicos permitindo apontar os MS2 multiplex e registrar os seus respectivos íons precursores. A consequência é a possibilidade de realização de buscas múltiplas para cada espectro multiplex; cada busca é restrita a um espaço de hipóteses determinado pela massa monoisotópica de um íon parental.

A comparação do YADA ao Xtract (Thermo Fisher, San José), mostrou que o primeiro é 700% mais rápido na velocidade de deconvolução. A análise de dados pós-processados por YADA a partir de experimento *middle-down* mostrou aumento do número de espectros identificados em quase 20%; isto se deve à maior sensibilidade na detecção de envelopes isotópicos, buscas (para identificações espectrais) com espaços de hipóteses reduzidos e consideração de espectros multiplex.   Os resultados do YADA encontram-se detalhados no **Artigo IX ("YADA: a tool for taking the most out of high resolution spectra")** do anexo.

3.5 Aumentar o número de peptídeos identificados e a confiabilidade de quantificação nos experimentos shotgun.

Este trabalho propõe nova maneira de aquisição de dados por espectrometria de massa aqui denominada como *extended data independent analysis* (XDIA). O XDIA combina MS1 de alta resolução com espectros multiplex de MS2 adquiridos por CAD e/ou ETD. Os dados são então processados por *software* desenvolvido neste trabalho denominado de *XDIA Processor,* que usa sub-rotinas do DA descrito no item anterior. Quando verificado em dados de proteômica *middle-down*, aumentamos em ~250% o número de espectros identificados e em ~30% o de peptídeos únicos quando comparados a dados adquiridos usando o tradicional *data-dependent aquisition* e a dissociação de íons por ETCaD. O aumento no número de espectros implica em maior confiança na quantificação de dados. O aumento no número de peptídeos únicos implica em maior "cobertura da sequência proteica". Os resultados do XDIA e maiores detalhes sobre o algoritmo e o método de aquisição de dados encontram-se descritos no **Artigo X ("XDIA: improving on the label free data-independent analysis")** do anexo. Vale ressaltar que estamos atualizando o XDIA *Processor* para também analisar *labeled data* (*e.g.*, SILAC (28)). Acreditamos que, devido ao grande melhoramento sobre a presente técnica estado-da-arte, o XDIA poderá levar a uma alteração no paradigma de como os experimentos de proteômica shotgun são realizados.

4 Discussão e conclusões

O estudo de proteômica *shotgun* permite analisar milhares de proteínas; assemelha-se a uma orquestra, cuja grandeza, não pode ser vislumbrada pela apreciação isolada de cada músico. Em contrapartida, cada músico é altamente especializado na função que lhe compete. Em "Getting closer to the whole picture", publicado recentemente na *Science* (29), Sauer U. *et. a*. relatam o impacto que esta classe de técnicas poderá trazer à clinica médica e na criação de novos fármacos. Resumidamente, os autores definem este ramo da ciência como a combinação de: modelagem computacional, matemática e experimentos quantitativos para descrever processos celulares. Os autores também enfatizam a carência de abordagens digitais para este propósito. A importância deste tipo de pesquisa reside compreender patologias ao nível molecular e fornecer ao médico painéis de alterações moleculares da patologia, possibilitando executar terapias individualizadas, acompanhar evolução e prognóstico através de marcadores específicos. Isto é exemplificado na observação de alterações "pré-malígnas" que possibilita ao clínico antecipar condutas úteis para cada indivíduo.

Nesta tese, apresentamos um ambiente computacional para analise de dados proteômicos que estende a capacidade das técnicas atuais. Introduzimos um algoritmo de deconvolução que serviu de base para a criação do XDIA. Este último, introduz um novo paradigma na forma de aquisição de dados de proteômica *shotgun* que permite maior precisão na quantificação de proteínas em misturas complexas aumentando, consideravelmente, o número de peptídeos identificados. Os resultados mostram, claramente, a superioridade do XDIA sobre a metodologia atual conhecida como *data-dependent analysis*. No presente trabalho foi também foi introduzido o CPM; este algoritmo possibilita a identificação de proteínas em experimentos do tipo *middle-down* executados em espectrômetros de baixa resolução. Na análise dos dados, foi introduzido o *PatternLab* com abordagens estatísticas para apontar proteínas diferencialmente expressas e auxiliar na interpretação dos resultados experimentais sob a luz do *Gene Ontology*. As ferramentas aqui introduzidas podem intercambiar dados, criando-se assim, um ambiente computacional para proteômica. O uso da computação aliada à biologia constitui novo ramo da ciência, a bioinformática, que nos capacita, cada vez mais, a vislumbrar como é regida a orquestra da vida.

5 Bibliografia

(1) Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science 2001 Feb 16;291(5507):1304-51.

(2) Hack CJ. Integrated transcriptome and proteome data: the challenges ahead. Brief Funct Genomic Proteomic 2004 Nov;3(3):212-9.

(3) Kramer DA. Commentary: Gene-environment interplay in the context of genetics, epigenetics, and gene expression. J Am Acad Child Adolesc Psychiatry 2005 Jan;44(1):19-27.

(4) Filipowicz W, Jaskiewicz L, Kolb FA, Pillai RS. Post-transcriptional gene silencing by siRNAs and miRNAs. Curr Opin Struct Biol 2005 Jun;15(3):331-41.

(5) Hodgetts R. Eukaryotic gene regulation by targeted chromatin re-modeling at dispersed, middle-repetitive sequence elements. Curr Opin Genet Dev 2004 Dec;14(6):680-5.

(6) Pandey RR, Ceribelli M, Singh PB, Ericsson J, Mantovani R, Kanduri C. NF-Y regulates the antisense promoter, bidirectional silencing, and differential epigenetic marks of the Kcnq1 imprinting control region. J Biol Chem 2004 Dec 10;279(50):52685-93.

(7) Soufla G, Porichis F, Sourvinos G, Vassilaros S, Spandidos DA. Transcriptional deregulation of VEGF, FGF2, TGF-beta1, 2, 3 and cognate receptors in breast tumorigenesis. Cancer Lett 2005 Jun 8.

(8) Chand HS, Ness SA, Kisiel W. Identification of a novel human tissue factor splice variant that is upregulated in tumor cells. Int J Cancer 2005 Oct 10.

(9) Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science 1989 Oct 6;246(4926):64-71.

(10) Karas, M., and Hillenkamp, F., 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. 1988 Oct 15;60(20):2299-301

(11) Liu H, Sadygov RG, Yates JR, III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 2004 Jul 15;76(14):4193-201.

(12) Coon JJ, Ueberheide B, Syka JE, Dryhurst DD, Ausio J, Shabanowitz J, et al. Protein identification using sequential ion/ion reactions and tandem mass spectrometry. Proc Natl Acad Sci U S A 2005 Jul 5;102(27):9463-8.

(13) Coon JJ. Collisions or electrons? Protein sequence analysis in the 21st century. Anal Chem 2009 May 1;81(9):3208-15.

(14) Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science 1989 Oct 6;246(4926):64-71.

(15) Barber, M., Bordoli, R. S., Sedgwick, R. D., and Tyler, A. N., 1981. Fast atom bombardment of solids (F.A.B.): a new ion source for mass spectrometry. J. Chem. Soc. Chem. Commun. 1981, 325-327

(16) Han B, Lenggoro IW, Choi M, Okuyama K. Measurement of cluster ions and residue nanoparticles from water samples with an electrospray/differential mobility analyzer. Anal Sci 2003 Jun;19(6):843-51.

(17) Mathieu E. Mémoire sur Le Mouvement Vibratoire d'une Membrane de forme Elliptique. Journal des Mathématiques Pures et Appliquées 1868 Jan 1;137-203.

(18) Makarov A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem 2000 Mar 15;72(6):1156-62.

(19) Liu H, Sadygov RG, Yates JR, III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 2004 Jul 15;76(14):4193-201.

(20) Eng JK, L.McCormack A, Yates, Yates JR, III. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J Am Soc Mass Spectrom 1994;5:976-89.

(21) Xu T, Venable JD, Park SK, Cociorva D, Lu B, Liao L, et al. ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. Mol Cell Proteomics 2006;5(S174).

(22) Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999 Dec;20(18):3551-67.

(23) Cociorva D, Tabb L, Yates JR. Validation of tandem mass spectrometry database search results using DTASelect. Curr Protoc Bioinformatics 2007 Jan;Chapter 13:Unit.

(24) Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res 1997 Oct;7(10):986-95.

(25) Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57:289-300.

(26) Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000 May;25(1):25-9.

(27) Rockwood AL, Haimi P. Efficient calculation of accurate masses of isotopic peaks. J Am Soc Mass Spectrom 2006 Mar;17(3):415-9.

(28) Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 2002 May;1(5):376-86.

(29) Sauer U, Heinemann M, Zamboni N. Genetics. Getting closer to the whole picture. Science 2007 Apr 27;316(5824):550-1.

ANEXO I

PubMed

U.S. National Library of Medicine
National Institutes of Health

Display Settings: Abstract

Genet Mol Res. 2008 Apr 15;7(2):342-56.

Identifying differences in protein expression levels by spectral counting and feature selection.

Carvalho PC, Hewel J, Barbosa VC, Yates JR 3rd.

Programa de Engenharia de Sistemas e Computação, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil. carvalhopc@cos.ufrj.br

Spectral counting is a strategy to quantify relative protein concentrations in pre-digested protein mixtures analyzed by liquid chromatography online with tandem mass spectrometry. In the present study, we used combinations of normalization and statistical (feature selection) methods on spectral counting data to verify whether we could pinpoint which and how many proteins were differentially expressed when comparing complex protein mixtures. These combinations were evaluated on real, but controlled, experiments (yeast lysates were spiked with protein markers at different concentrations to simulate differences), which were therefore verifiable. The following normalization methods were applied: total signal, Z-normalization, hybrid normalization, and log preprocessing. The feature selection methods were: the Golub index, the Student t-test, a strategy based on the weighting used in a forward-support vector machine (SVM-F) model, and SVM recursive feature elimination. The results showed that Z-normalization combined with SVM-F correctly identified which and how many protein markers were added to the yeast lysates for all different concentrations. The software we used is available at http://pcarvalho.com/patternlab.

PMID: 18551400 [PubMed - indexed for MEDLINE]

PMCID: PMC2703009

Publication Types, MeSH Terms, Substances, Grant Support

LinkOut - more resources

You are here: NCBI > Literature > PubMed

Write to the Help Desk

ANEXO II

BMC Bioinformatics

BioMed Central

Software

Open Access

PatternLab for proteomics: a tool for differential shotgun proteomics

Paulo C Carvalho*[1], Juliana SG Fischer[2], Emily I Chen[3], John R Yates III[3] and Valmir C Barbosa[1]

Address: [1]Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, [2]Laboratory for Protein Chemistry, Chemistry Institute, and the Rio de Janeiro Proteomic Network, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil and [3]Biological Mass Spectrometry Laboratory, The Scripps Research Institute, La Jolla, California, USA

Email: Paulo C Carvalho* - carvalhopc@cos.ufrj.br; Juliana SG Fischer - juli_f@iq.ufrj.br; Emily I Chen - emily@pharm.stonybrook.edu; John R Yates - jyates@scripps.edu; Valmir C Barbosa - valmir@cos.ufrj.br

* Corresponding author

Published: 21 July 2008

BMC Bioinformatics 2008, **9**:316 doi:10.1186/1471-2105-9-316

Received: 25 December 2007

Accepted: 21 July 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/316

Abstract

Background: A goal of proteomics is to distinguish between states of a biological system by identifying protein expression differences. Liu *et al.* demonstrated a method to perform semi-relative protein quantitation in shotgun proteomics data by correlating the number of tandem mass spectra obtained for each protein, or "spectral count", with its abundance in a mixture; however, two issues have remained open: how to normalize spectral counting data and how to efficiently pinpoint differences between profiles. Moreover, Chen *et al.* recently showed how to increase the number of identified proteins in shotgun proteomics by analyzing samples with different MS-compatible detergents while performing proteolytic digestion. The latter introduced new challenges as seen from the data analysis perspective, since replicate readings are not acquired.

Results: To address the open issues above, we present a program termed PatternLab for proteomics. This program implements existing strategies and adds two new methods to pinpoint differences in protein profiles. The first method, ACFold, addresses experiments with less than three replicates from each state or having assays acquired by different protocols as described by Chen *et al.* ACFold uses a combined criterion based on expression fold changes, the AC test, and the false-discovery rate, and can supply a "bird's-eye view" of differentially expressed proteins. The other method addresses experimental designs having multiple readings from each state and is referred to as nSVM (natural support vector machine) because of its roots in evolutionary computing and in statistical learning theory. Our observations suggest that nSVM's niche comprises projects that select a minimum set of proteins for classification purposes; for example, the development of an early detection kit for a given pathology. We demonstrate the effectiveness of each method on experimental data and confront them with existing strategies.

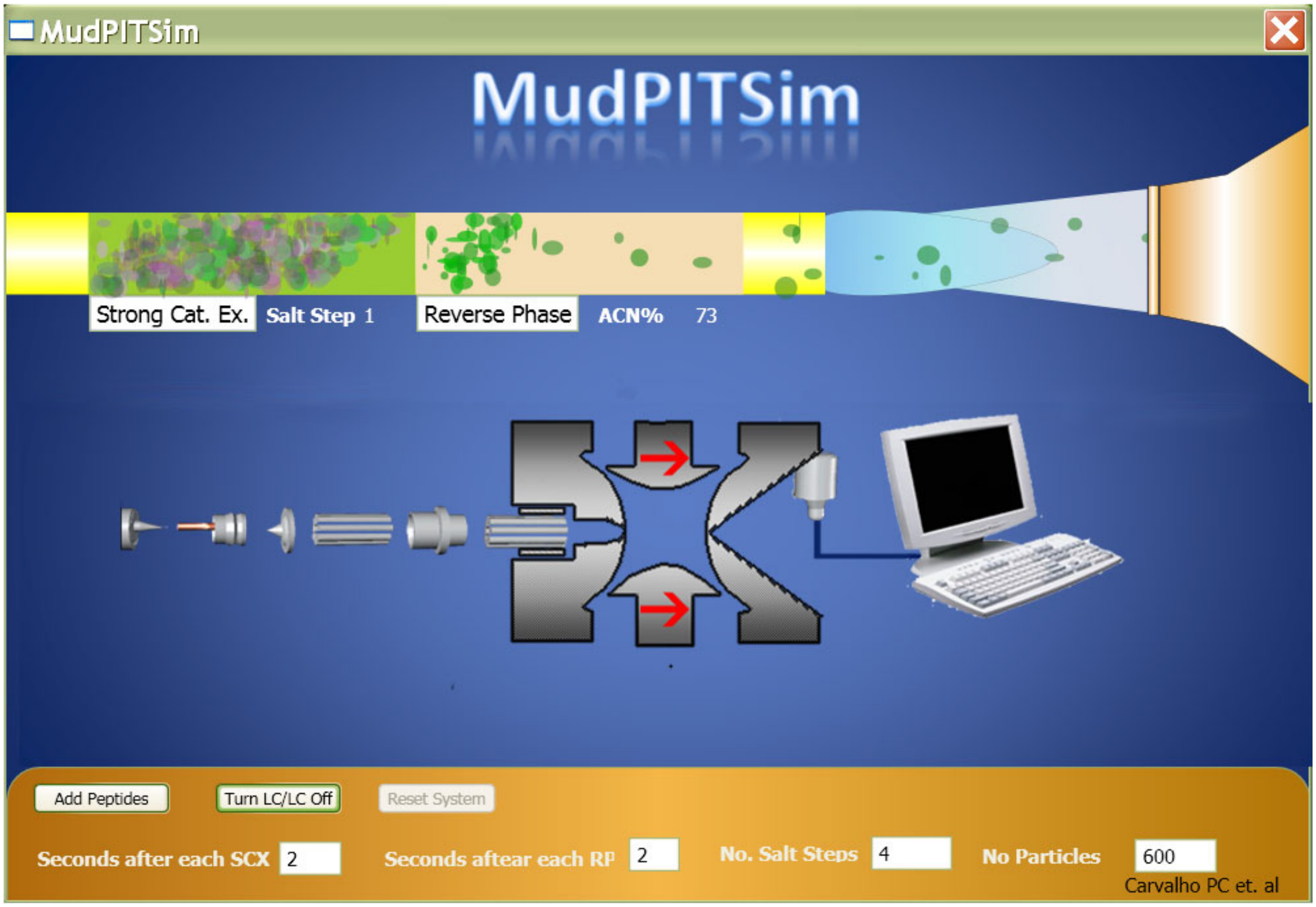
Conclusion: PatternLab offers an easy and unified access to a variety of feature selection and normalization strategies, each having its own niche. Additionally, graphing tools are available to aid in the analysis of high throughput experimental data. PatternLab is available at http://pcarvalho.com/patternlab.

Background

A goal of proteomics is to distinguish between states of a biological system by identifying protein expression differences [1]. Shotgun proteomics is a large-scale strategy for protein identification in complex mixtures that involves pre-digestion of intact proteins followed by peptide separation, fragmentation in a mass spectrometer, and database search. Its name is derived from DNA shotgun sequencing, which in turn follows the analogy of a shotgun's quasi-random firing pattern and dispersion to ensure the target is hit.

Multi-dimensional Protein Identification Technology (MudPIT) is a shotgun proteomics technique capable of identifying thousands of proteins in proteolytically digested complex mixtures [2,3]. MudPIT separates peptides according to two independent physicochemical properties using two-dimensional liquid chromatography (LC/LC) online with the ion source of a mass spectrometer. This separation relies on columns of strong cation exchange (SCX) and reversed phase (RP) material, back to back, inside fused silica capillaries. The chromatography proceeds in cycles, each of which consists of increasing salt concentration to "bump" peptides off the SCX followed by a hydrophobic gradient to progressively elute peptides from the RP into the ion source. This process identifies mixture components by tandem mass spectrometry (MS/MS). For didactic purposes, a simplified and interactive MudPIT simulator is available at the project's web site; its interface is described in Figure 1.

Computational approaches for LC/MS-based differential proteomics usually involve *in silico* chromatogram alignment followed by pattern recognition strategies [4]. How-



Salt Step 1

ACN% 73

Seconds after each SCX 2

Seconds aftear each RP 2

No. Salt Steps 4

No Particles 600

Figure 1
MudPIT simulator. The image displays the graphical user interface of the MudPIT simulator available on the project's website for didactic purposes. The simulator allows one to specify MudPIT parameters and then see the two-dimensional liquid chromatography simulation proceed on the fly. This is a simplification of reality; therefore, the timescale and many other features are not faithful representations. The green and pinkish structures in the upper part of the simulator represent the strong cation exchange and the reverse phase material packed in the capillary (yellow structure). The semi-conical structure represents the mass spectrometer nozzle (entrance) and the structure below is an X-Ray of a quadrupole ion trap.

ever, because of the more complex nature of MudPIT's LC/LC method and the alternating acquisition of mass spectra and tandem mass spectra, chromatographic alignment is more complicated than for LC/MS data. A milestone that eventually allowed differential MudPIT analysis was set with the development of alternative protein quantitation methods that use features from the tandem mass spectra (e.g., peptide hits, protein sequence coverage, spectral counts) as surrogate measures of protein abundance [2,5-7]. An important step was taken when Liu *et al.* demonstrated that the number of tandem mass spectra obtained for each protein, or "spectral count", correlates linearly with protein abundance in a mixture for two orders of magnitude [8]. These advances allowed LC/LC/MS/MS to produce semi-quantitative data on mixtures; however, two issues have remained open: how to normalize spectral count data for profile comparisons and how to statistically identify *bona fide* differences between samples (feature selection). Heretofore, differential proteomics by MudPIT spectral counting has relied on repeating assays to increase the number of identified proteins, improve protein coverage, and enable traditional statistical methods to pinpoint differences between biological states. Studies have shown Student's t-test, Fisher's exact test, and the G-test to be trustworthy for composing putative differential marker tallies when three or more replicates are available [9].

Recently, Chen *et al.* increased peptide and protein identifications in complex protein mixtures by re-analyzing samples digested in the presence of different MS-compatible detergents [10]. Moreover, the improved proteolytic digestion protocols potentially increased identification of less abundant proteins. However, the experimental design described by Chen *et al.* introduced additional data analysis challenges, since replicate readings are not acquired.

The contributions by Liu *et al.* and Chen *et al.* serve as foundations for this work. Here, we introduce a simple to use, yet efficient and panoptic, software for differential shotgun proteomics that addresses the data analysis issues of the experimental designs mentioned above. Our software, PatternLab for proteomics, or just PatternLab as referred to throughout, achieves its goal by featuring two new data analysis methods in addition to other widely adopted statistical approaches. The first method, ACFold, addresses experiments with less than three replicates from each state (class) or having data acquired by different protocols, as described by Chen *et al.* ACFold uses a combined criterion based on expression fold changes and the AC test [11]; its importance is demonstrated here with experimental data. The other method addresses experimental designs that comprise multiple replicates from each state and is referred to as nSVM (natural support vector machine) because of its roots in evolutionary computing and statistical learning theory [12]. We benchmarked nSVM against the widely adopted Student's t-test over a spiked marker dataset and identified its niche. A detailed description of ACFold, nSVM, and PatternLab's overall architecture is given, and critical issues of each method and how they were addressed are provided in the Implementation section.

Implementation

PatternLab's current version is optimized for LC/LC/MS/MS data using spectral counts. Its architecture comprises four core modules (parsing MS data, data normalization, feature selection, and analysis). These modules can be operated programmatically or through the graphical user interface (GUI) that also provides specialized graphing tools to aid interpretation. Details of each module and a walkthrough of PatternLab, including the two new feature selection procedures ACFold and nSVM, are described below.

Parser module

Let "project" refer to one's experimental data from all MudPIT assays of all biological samples from both control and case states. PatternLab relies on the parser module to translate a project's MS data into an index file and a sparse matrix file. The index file lists all identified proteins within the project and assigns each one a unique Protein IDentification (PID) integer. As for the sparse matrix, each row follows the schema: $\langle$class label$\rangle\langle$PID$\rangle$:$\langle$value$\rangle$...$\langle$PID$\rangle$:$\langle$value$\rangle$. In the latter, $\langle$class label$\rangle \in \{-1, +1\}$ is used to identify a biological state (e.g., +1 for control and -1 for case); $\langle$PID$\rangle$ and $\langle$value$\rangle$ correspond, respectively, to a protein identification index in the project's index file and to the spectral count verified for that protein during the corresponding MudPIT analysis. So, for example, the row "+1 1:3 2:5 3:6" specifies an analysis from the positive class having spectral count values of 3, 5, and 6 for PIDs 1, 2, and 3, respectively, all other PIDs having value 0.

There are various softwares that identify proteins by matching tandem mass spectra according to a database of peptide sequences, such as SEQUEST [13] and MASCOT [14]. The current parser can address both SEQUEST followed by DTASelect [15] and MASCOT having results exported to the DTASelect format. To use the parser, one should place the DTASelect results from the control and case analyses in different folders and then simply indicate their paths in the GUI.

Normalization module

One or more normalization methods can be applied to the sparse matrix. PatternLab currently implements: ln (natural logarithm), Z [16], Total Signal, Maximum Signal, and Row Sigma. The ln normalization is obtained by taking the natural logarithm of every value and aims at increasing the signal of the PIDs with low spectral counts with respect to the more abundant PIDs. The Z normalization is achieved by subtracting from each original value the mean of all values of the corresponding PID and dividing the result by the standard deviation of all values from the same PID; the mean then becomes 0 and the standard deviation 1. The Total Signal normalization is achieved by dividing each value by the sum of all values in the respective row. The Maximum Signal normalization is obtained by dividing each value by the largest value in its row; an underlying assumption is that, in each MudPIT analysis, peptide identifications were obtained at or near the capacity of the tandem MS instrument. The Row Sigma normalization is achieved by calculating the mean and standard deviation of all values in a row and then dividing each value by the mean plus three standard deviations. The latter is introduced in this work as a variation of the Maximum Signal normalization that better handles assays that obtained an exceedingly high maximum value for a protein; further advantages are addressed in the Results and discussion section.

ACFold feature selection

The ACFold analysis introduced in this paper is intended to evaluate data from projects having less than three replicate assays per class or assays obtained using different mass spectrometry protocols as described by Chen *et al.* [10]. The ACFold analysis takes advantage of two accepted criteria in proteomics to pinpoint differences between samples: the generalized AC Test [11] and expression fold changes [9]. The algorithm first parses the project's data as described in the Parser module section. The sparse matrix is then compressed into two rows, one representing each class. The new rows' values for each PID are obtained by averaging the original values of the corresponding PIDs within their classes. But given the nature of the experiment at hand, it is likely that low-probability events (such as PIDs that obtained a spectral count of 1 in only one assay of one of the two classes) will not always be observed; this would result in a calculated average of 0 and imply a probability of 0 that is not justifiable by evidence according to Cromwell's rule. To avoid the zero-frequency problem [17] and make fold-change calculations possible, a pseudo spectral count of 1 is then added to each PID value of the two resulting rows, including the unobserved PIDs, following the process known as Laplace's rule of succession. PatternLab then calculates the AC test probabilities and the expression fold changes according to one of the user-specified normalization methods: Total Signal, Row Sigma, or None.

Finally, a false-discovery rate (FDR) is estimated by the Benjamini-Hochberg procedure [18] for a given fold-change cutoff. Let m be the number of identified proteins minus the number of proteins that failed to pass the fold-change cutoff test. For $1 \leq i \leq m$, let H_i^0 be the (null) hypothesis that the ith protein is not differentially expressed, and p_i its p-value. Assuming $p_1 \leq p_2 \leq \ldots \leq p_m$ (ties are broken so that no lower-fold protein ranks ahead of another having the same p-value), and α the minimum FDR at which a test can be called significant, let

$$k = \max\left\{ i : p_i \leq \frac{i}{m}\alpha \right\}.$$

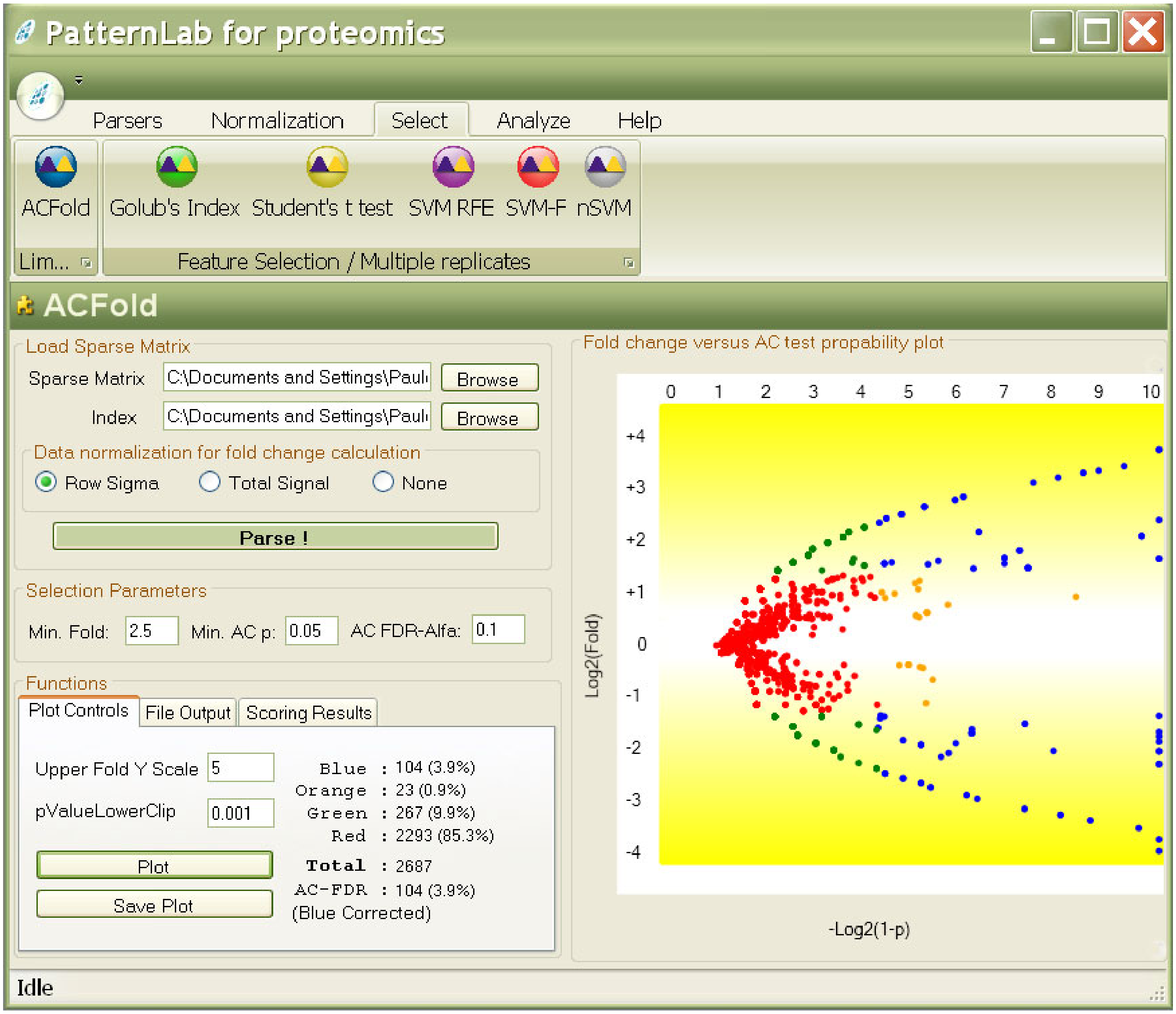
The null hypotheses H_k^0 are then rejected (i.e., the corresponding proteins are declared differentially expressed). If no such i exists, no hypothesis is rejected.

The user can define stringency levels aided by a distribution plot and supplementary information offered by the GUI indicators. The stringency is performed by specifying a minimum fold-change cutoff, an AC test p-value cutoff, and the FDR α. We refer to Figure 2 to demonstrate how the results are presented. Lastly, the final report can be exported to text.

nSVM feature selection

nSVM is a feature selection algorithm introduced in this work and used here to pinpoint differences in protein expression profiles when multiple replicates of each class are available. The algorithm begins by parsing the project's data as described in the Parser module section. nSVM then uses the structural risk minimization (SRM) principle from statistical learning theory [12] to drive the convergence of a genetic algorithm (GA). Briefly, a GA is a stochastic optimization technique inspired in evolutionary biology which imitates inheritance, selection, crossover, and mutation to evolve a population of abstract genomes (individuals) [19]. Each individual represents a candidate solution (set of differentially expressed PIDs) and is coded as an array of bits (1 or 0); the nth bit value hypothesizes that either the protein whose PID value is n is differentially expressed (1) or not (0). The general aim of a GA is to evolve an initial population of randomly generated individuals so that, after a number of generations, the solution will be encoded in the genome of the historically fittest individual.

The GA works by generating successive populations on the premise that the average individual fitness (quality of the solution) will increase for each new population. Each new solution from nSVM is produced by first selecting parents according to their quadratically normalized fitnesses. Formally, let S denote the set $\{i_0, i_2, \ldots, i_{n-1}\}$ of individuals, ordered by nondecreasing fitness. Let j and k be two randomly chosen numbers in the range from 0 to n^2- 1. The



Sparse Matrix C:\Documents and Settings\Paul

Index C:\Documents and Settings\Paul

● Row Sigma ○ Total Signal ○ None

Min. Fold: 2.5 Min. AC p: 0.05 AC FDR-Alfa: 0.1

Upper Fold Y Scale 5

pValueLowerClip 0.001

```
  Blue  : 104 (3.9%)
Orange  : 23 (0.9%)
 Green  : 267 (9.9%)
   Red  : 2293 (85.3%)

 Total  : 2687
AC-FDR  : 104 (3.9%)
(Blue Corrected)
```

Figure 2

ACFold's graphical user interface. The interface above displays results from real experimental data. The plot on the right shows the distribution of the identified proteins according to $\log_2$(fold change) on the ordinate (y) and – $\log_2$(1- (AC test *p*-value)) on the abscissa (x). The plot tab indicates that 104 proteins (blue dots) were differentially expressed because they satisfied both the AC test and fold-change cutoffs specified by the user. 23 proteins (orange dots) did not meet the fold-change cutoff but were indicated as statistically differentially expressed, therefore deserving a second look. 267 proteins (green dots) met the fold-change cutoff; however, the AC test indicated that this happened by chance. 2293 proteins (red dots) were pinpointed as not differentially expressed between classes because they failed both the AC test and the fold-change cutoffs. The GUI also lists an AC FDR indicating that all blue dots satisfy the established user-selected FDR of 0.1.

two individuals chosen to mate will be the ones in *S* indexed by the greatest integers no greater than the square roots of *j* and *k* (i.e., the square roots' "floors"); if the same individual is chosen twice, this process is repeated. Clearly, fitter individuals have significantly higher chances of being selected. During the mating process, a uniformly random crossover operator is used, so the single offspring receives each gene (bit value) from either one of its parents with equal chances. The GA then performs mutations on the newly produced offspring according to a user-specified mutation index. For example, a mutation index of 2 allows the GA to perform up to two mutations in the offspring's genome. The mutation is performed by switching the values of randomly chosen bits with a bias towards mutating them to 0 (specifically, a 60% chance for 0 and 40% for 1). We recall that 0 and 1 represent excluding or including a feature, respectively. This bias accelerates the GA in finding solutions with fewer features. In addition, a fine-tuning parameter termed mutInd1After can be set through the GUI. This parameter stands for "Mutation Index 1 After", so after the algorithm has reduced the initial set of candidate proteins to a number below the one the parameter specifies, the mutation index is reduced to 1. This allows the GA to search within the remaining combinations with a lower probability of making great shifts away from the local optimum it is approaching. The processes of mating, crossover, and mutation are repeated until a population of the same size as the initial one is formed for use in the next iteration of the algorithm. The user can also configure the GA to allow "elitism", permitting a specified fraction of the fittest individuals to continue on to the new population. The algorithm terminates when a user-specified number of generations has elapsed without the appearance of an individual that is fitter than the fittest found so far.

Fitness evaluation is certainly one of the most important aspects of a GA. As far as we know, this is the first time a GA takes advantage of the SRM principle [12] to drive its convergence. Briefly, the SRM principle allows the evaluation of how well data points are separated in a feature space by a classification function, according to an empirical error measure on known examples and an upper bound on the function's error when generalizing for unknown examples [12]. The SRM principle is the basis of the SVM pattern recognition method, which searches for a classification function with the "best" trade-off between empirical error and worst-case generalization error. The upper bound on the generalization error grows monotonically with the machine's so-called VC dimension, so lower VC dimensions are preferred. Additionally, another upper bound on the generalization error depends on the machine's number of support vectors in a way that a small number of such vectors is also preferred [12]. We use this other bound as well. In the remainder of this section we refer to each row of the sparse matrix as an input vector.

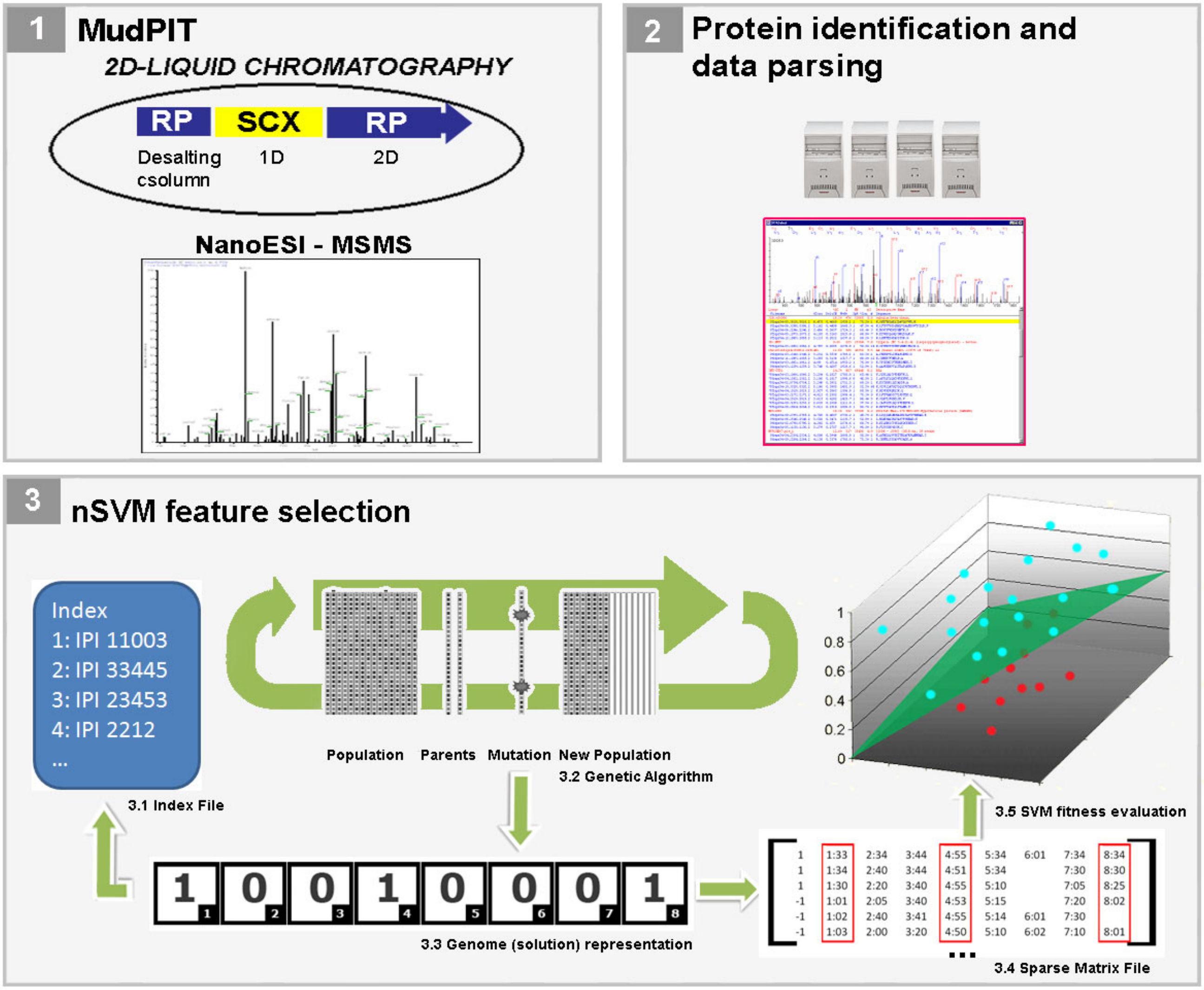
Each individual's fitness is evaluated by how well the input vectors are "separated" in the feature space defined by the individual. First, the input vectors are mapped onto the feature space taking into consideration only the proteins whose PIDs have value 1 in the individual. Secondly, an SVM model is generated and the empirical error is evaluated by the leave-one-out approach [20]. The VC dimension, the number of support vectors, and the number of bits having value 1 in the individual are also recorded. Finally, the fitness score for an individual is calculated as

$$\mathit{fitnessScore} = C_1 LOO + C_2\left(1-\frac{1}{h}\right) + C_3\left(1-\frac{1}{nSV}\right) + C_4 nG$$

where *LOO* is the SVM leave-one-out error, *h* is the VC dimension, *nSV* is the number of support vectors, *nG* is the number of bits with value equal to 1, and C_1 through C_4 are user-specified constants having default values set to 100, 100, 10, and 0.1, respectively. Clearly, the lower the score, the fitter the individual. We note that the first three parameters are calculated using SVM *light* [21]. Figure 3 summarizes the nSVM process up to this point.

nSVM relies on the island model to keep population diversity and to better address the issue of a large search space. This approach works with a user-specified number of populations that evolve independently. Individuals will migrate, from time to time, according to a user-specified time parameter. The migration proceeds as follows. First the GA randomly chooses two populations from its pool and pauses their computations after the fitness evaluation step. Secondly, a random number is picked (conforming to a user-specified upper bound) to indicate the number of individuals to be exchanged between the two populations. Thirdly, individuals are selected (as for mating, described above) and are exchanged between the populations. Finally, the GA continues to evolve both populations from where they were stopped. PatternLab takes advantage of the recent multi-core processors by having each population "live" in a different computing thread. Thus, a computer with a certain number of cores can manage as many populations concurrently without sacrificing performance.

The features for the final classification model are selected by executing nSVM multiple times (e.g., 20). For every nSVM execution, each time the fittest individual is replaced its genomic information is saved in a text file (history file). After multiple nSVM executions, several history files are available and a ranking of the features can be established according to the frequency of occurrence of each PID in the history files. Furthermore, a number of minimal discriminative features can be estimated by generating a two-column list having PIDs ordered by their ranks in the first column and their achieved frequencies in the second. The set of discriminative features is then estimated by locating, in this list, the two consecutive rows that present the greatest difference in frequency values. The number of features is then computed by counting how many features have scores above or equal to this gap's upper limit.



1	1:33	2:34	3:44	4:55	5:34	6:01	7:34	8:34
1	1:34	2:40	3:44	4:51	5:34		7:30	8:30
1	1:30	2:20	3:40	4:55	5:10		7:05	8:25
-1	1:01	2:05	3:40	4:53	5:15		7:20	8:02
-1	1:02	2:40	3:41	4:55	5:14	6:01	7:30	
-1	1:03	2:00	3:20	4:50	5:10	6:02	7:10	8:01

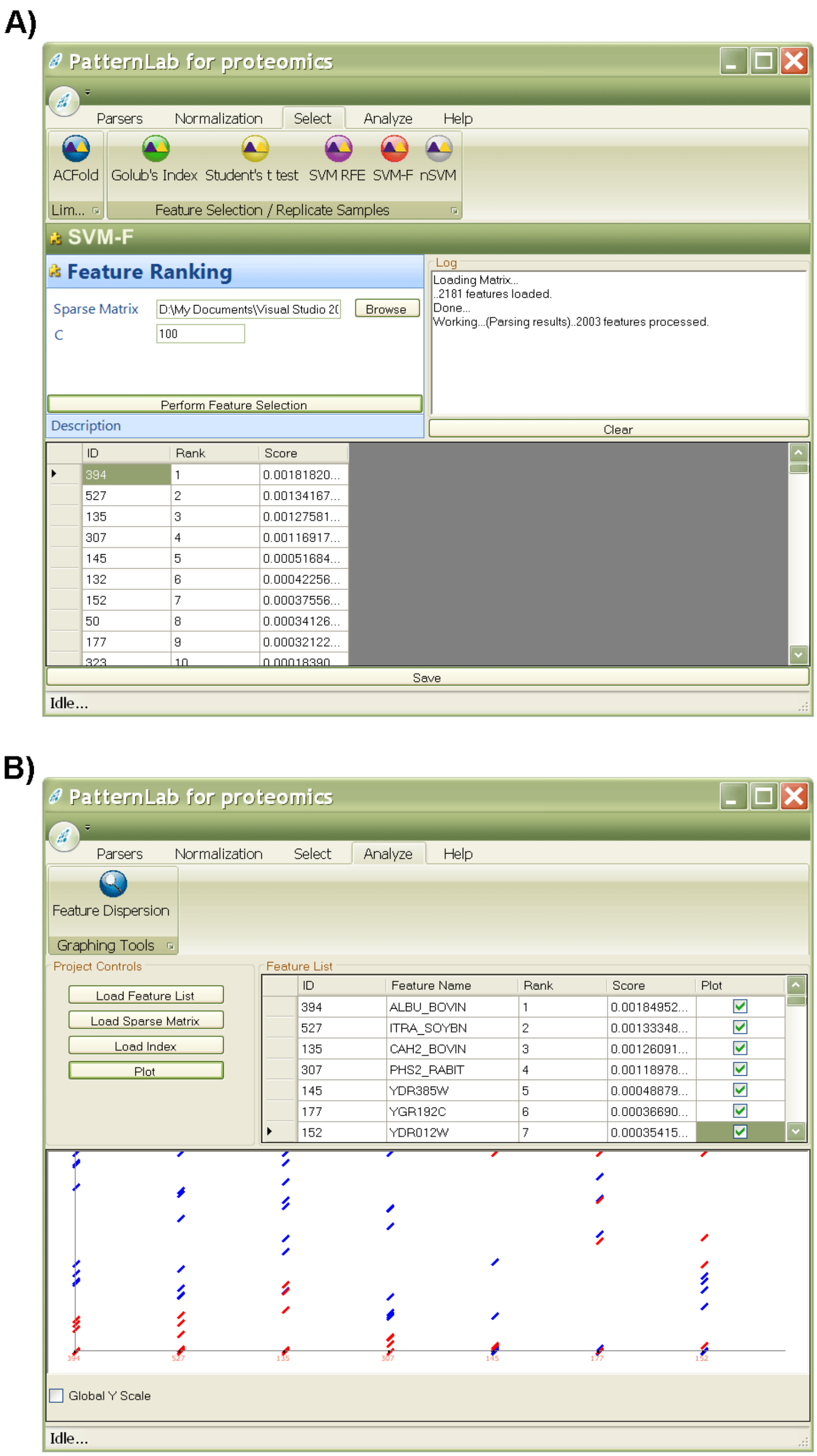
Figure 3
nSVM's workflow. MudPIT is applied to acquire mass spectrometry data from a biological system in different states (1). The data are subsequently identified by SEQUEST and filtered by DTASelect (2). nSVM is applied to pinpoint differences in the protein expression profiles by using a GA (3.2). Each individual's genome is an array of bits (3.3) that corresponds to a set of proteins (3.1 and 3.2) that will be selected from the dataset (3.4) to be evaluated as a solution (3.5) according to their spectral counts.

Other available statistical inference methods and the result analyzer module

PatternLab offers several additional feature selection methods that are widely adopted by the proteomics community. These methods include: SVM recursive feature elimination (SVM-RFE) [22], forward SVM (the weighting used in the first step of SVM-RFE), Golub's index [23], and Student's t-test [11]. It is beyond the scope of this manuscript to detail these methods since they are well documented in the literature. Figure 4 exemplifies PatternLab's GUI to access the feature selection methods and a result analyzer. In a future version of PatternLab, we intend to add new components to the result analysis module. Figure 5 exemplifies nSVM's interface.

Results and discussion

Two main issues characterize feature selection challenges in bioinformatics: the large input dimensionality and limitations in the dataset size. To deal with these problems, various feature selection techniques have been designed by experts from the machine learning and data mining fields. The philosophy behind PatternLab is that there is



Sparse Matrix: D:\My Documents\Visual Studio 2(

C: 100

```
Loading Matrix...
..2181 features loaded.
Done...
Working...(Parsing results)..2003 features processed.
```

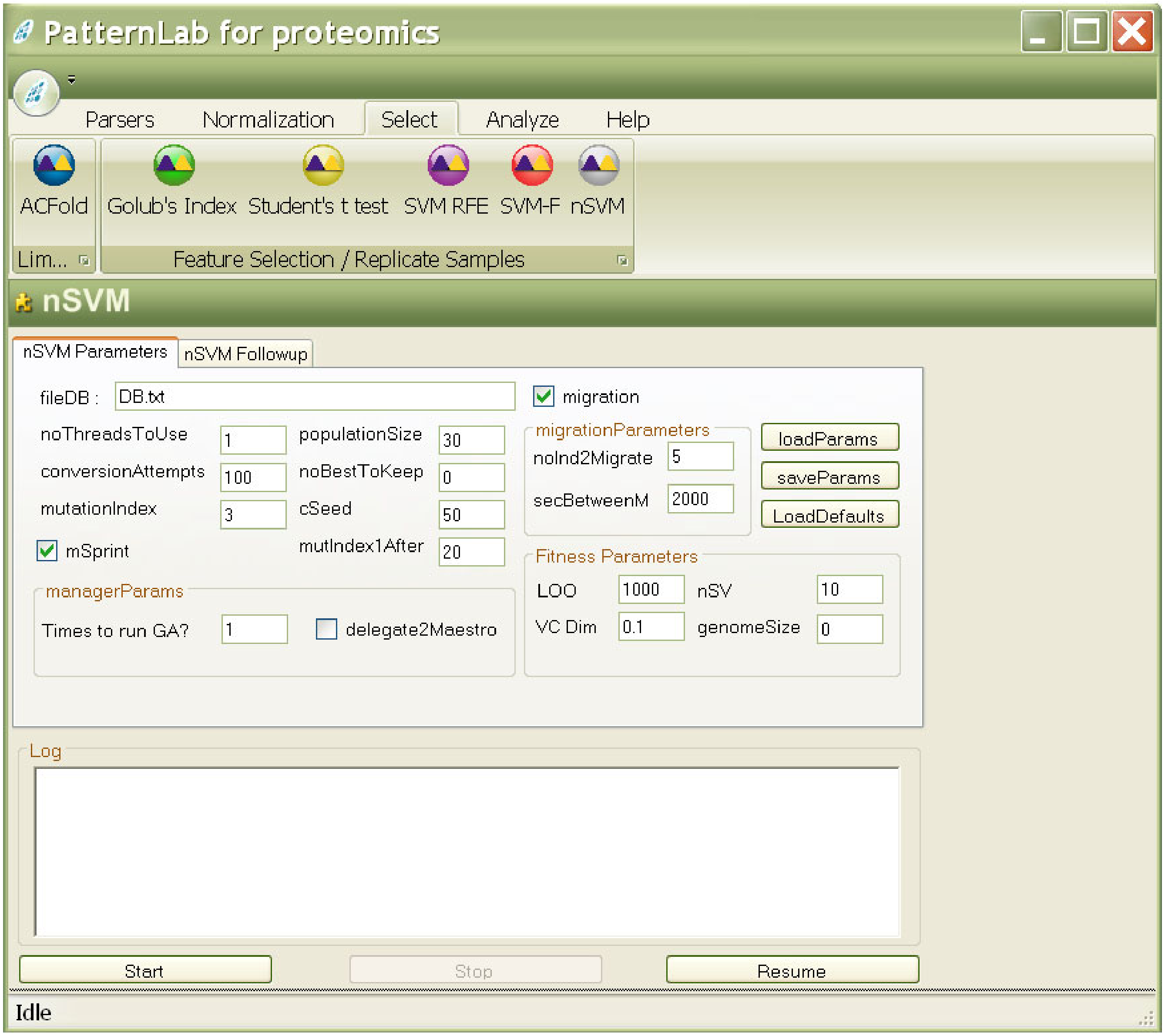
ID	Rank	Score
394	1	0.00181820...
527	2	0.00134167...
135	3	0.00127581...
307	4	0.00116917...
145	5	0.00051684...
132	6	0.00042256...
152	7	0.00037556...
50	8	0.00034126...
177	9	0.00032122...
323	10	0.00018390

ID	Feature Name	Rank	Score	Plot
394	ALBU_BOVIN	1	0.00184952...	☑
527	ITRA_SOYBN	2	0.00133348...	☑
135	CAH2_BOVIN	3	0.00126091...	☑
307	PHS2_RABIT	4	0.00118978...	☑
145	YDR385W	5	0.00048879...	☑
177	YGR192C	6	0.00036690...	☑
152	YDR012W	7	0.00035415...	☑

□ Global Y Scale

Figure 4

Replicate experiment analyzer's graphical user interface. This graphical user interface offers various normalization and feature selection methods (A). After applying the methods, the user can view the features ranked according to their scores. The expression from the selected feature can be graphed in the result analyzer (B).



fileDB : DB.txt

☑ migration

noThreadsToUse 1

populationSize 30

conversionAttempts 100

noBestToKeep 0

mutationIndex 3

cSeed 50

☑ mSprint

mutIndex1After 20

managerParams

Times to run GA? 1

□ delegate2Maestro

migrationParameters

noInd2Migrate 5

secBetweenM 2000

Fitness Parameters

LOO 1000

nSV 10

VC Dim 0.1

genomeSize 0

Figure 5
nSVM's graphical user interface. Every aspect of nSVM's GA can be customized in its graphical user interface or programmatically. A detailed explanation of each parameter can be obtained at the project's website.

no single, universally optimal feature selection technique [24]; additionally, the existence of more than one subset of features that discriminates the data equally well [25] should be considered. We believe that each feature selection strategy has its own niche, so it is important to know its idiosyncrasies, when to effectively apply it, and also to be aware of its limitations. For example, while the output provided by univariate feature rankings can be more intuitively grasped because they analyze each feature independently, protein subgroups that could possibly interact can only be detected through multivariate techniques (but requiring far more effort).

Row Sigma normalization

Methods with ease of interpretation tend to be more readily accepted, which is in line with the possibility of intuitive interpretation that is one of the goals of the Row Sigma normalization strategy introduced in this work. This strategy joins the robustness of the Total Signal normalization (by using a measure that considers the entire profile through the sum of all values) with the ease of interpretation of the Maximum Signal normalization (by characterizing the proteins as percentages of an estimated most abundant protein). This is achieved by dividing the expression value of each protein by the mean plus three times the standard deviation taken inside its row in the sparse matrix. When compared to the Maximum Signal normalization, which relates all proteins solely to the most abundant one, Row Sigma normalization is seen to help avoid misleading conclusions that might be reached should the most abundant protein have a large variance associated with it.

Suggestion of when to apply ACFold

ACFold combines the AC test with fold changes to pinpoint differentially expressed proteins between classes; this is important because conclusions drawn only from fold changes could be equivocated. For example, suppose spectral counts of 3 (6) and 30 (60) were observed for protein x (y) during the control (case) assay. Both x and y have a twofold up-regulation from the control to the case assay, but it is much likelier that the fold change for protein x happened by chance. The conditional probability of finding a spectral count of x_2 in biological state 2 given that a spectral count of x_1 was found in biological state 1 can be estimated by the AC test.

The AC test outputs a *p*-value related to testing a single hypothesis. In large-scale proteomic strategies, such as MudPIT, thousands of hypotheses are tested simultaneously, requiring an appropriate error rate control instead of relying solely on *p*-values. The most well-known strategy to deal with multiple hypotheses is the Bonferroni, but it is in some ways too conservative and this has led to the proposal of new ones [18,26]. The solution we propose to this massive multiple-hypothesis test problem is to analyze the data from an FDR perspective instead of that of *p*-values. The FDR is defined as the expected proportion of false positives among the results declared significant [18]. For example, by specifying an FDR of 0.1, it is expected that no more than 10% of the results declared significant (*p*-value ≤ cutoff) be false positives. This control, however, cannot be obtained with the *p*-value alone.

Label-free shotgun proteomics currently uses a random sampling process to estimate the relative quantitation of thousands of proteins. For this reason, determining the true number of differentially expressed proteins, which would require precise quantitation instead, has remained an open challenge. Due to the lack of a training set with known answers, our approach relies on a theoretical FDR estimator to cope with imprecise quantitation. One strategy to evaluate the effectiveness of FDR approaches is to spike protein markers with known concentrations into complex protein mixtures (e.g., lysates) to perform real, but controlled, experiments, which are therefore verifiable. For example, Zhang *et al.* [9] compared replicate LC-MS/MS assays of the *S. cerevisiae* lysate plus six protein markers (accounting for 1.25% and 2.5%, thus twofold, of the total protein content in the different classes) and observed that the true FDR of the AC test, when comparing two assays, varied from ~1% to ~13%, depending on the marker. Furthermore, the authors concluded that Fisher's exact test, the *G*-test, and the AC test all "give reasonable false positive rates even with limited sampling numbers from a single replicate." In view of these results, our choice to rely on a theoretical estimator instead of physical measurements seems justified.

Figure 2 exemplifies the ACFold analysis on experimental data. The aim was to identify as many proteins as possible in a glioblastoma cell culture when exposed (or not) to a chemotherapeutic agent during 1.5 h [27]. To maximize protein identification, different MS-compatible detergents were used during each MudPIT assay as described by Chen *et al.* [10]. Experiments that do not acquire replicate readings such as the latter or that have very few readings for each class (one or two) fall within ACFold's niche. As shown in Figure 2, 2687 proteins were identified, of which 104 were pinpointed as being differentially expressed according to ACFold (using a minimum fold-change cutoff of 2.5, an AC test *p*-value of 0.05, and an FDR of 0.1). The fold-change cutoff was empirically selected to maximize the number of hypotheses approved by the FDR criterion. For example, by specifying a cutoff of 2.0, the number of differentially expressed proteins according to the AC test was raised to 105; however, because of a great increase in the number of hypotheses tested, the FDR indicated only 88 as differentially expressed. When a cutoff of 3.0 was specified, both the number of differentially expressed proteins according to the AC test and the FDR were reduced to 85. We note, finally, that the FDR estimation is conservative for massively-multiple hypothesis testing [28] and that a high stringency on false positives can imply an increase in the number of false negatives; whence the choice of our FDR of 0.1.

Suggestion of when to apply nSVM

Our observations suggest that nSVM's niche comprises projects targeting the selection of a minimum set of proteins (features) that nevertheless allows the highest rate of correctness to be achieved on unseen samples in classification problems. This selection entails the solution of the difficult bioinformatics combinatorial problem of choosing one out of the 2^n sets into which n identified proteins can be combined. Two widely adopted classes of methodologies to solve this problem are the filter and the wrapper approaches. Briefly, the filter approach relies on a probabilistic method to eliminate or rank features, similarly for example to our use of the t-test. However, according to

Cover and Van Campenhout [29], no ordering of error probabilities is guaranteed to produce the optimal feature subset or subsets. Moreover, feature sets can be algorithm-dependent to achieve good results. On the other hand, wrapper methods handle the problem by relying on the classification algorithm during the feature selection process, but the algorithm becomes more prone to overfitting.

nSVM is a wrapper feature selection approach that couples a nature-inspired optimization technique (a GA) with the state-of-the-art classifier (an SVM) to address the overfitting problem. This hybrid approach is justifiable because, even though SVMs efficiently generalize on noisy datasets, they have no internal feature relevance evaluation; therefore, noisy features can degrade their performance. Consequently, feature selection plays a key role prior to SVM classification, especially for complex datasets as in shotgun proteomics. Our GA is a stochastic heuristic that deals with massive resampling to handle the idiosyncrasies of a dataset as related to a classifier to avoid overfitting. Additionally, the feature sets selected by our GA are optimized for classification by an SVM because our GA's fitness function considers the same principles that drive the SVM classifier (i.e., the empirical error, the VC dimension, and the number of support vectors). Accordingly, we showed that nSVM efficiently dealt with the overfitting problem on a high-dimensional and noisy dataset by correctly pinpointing the relevant features (spiked proteins) and outperforming the t-test filter approach, as described below.

We demonstrate nSVM's niche using data from a real (yeast lysate replicates), yet controlled, experiment (protein markers were spiked to simulate differences), which is therefore verifiable. The dataset was obtained from four aliquots of 400 μg of a soluble yeast total cell lysate that were mixed with Bio-Rad SDS-PAGE low range weight standards containing phosphorylase b (PHS2), serum albumin (ALB), carbonic anhydrase (CAH), and trypsin inhibitor (ITRA) at relative levels of 25%, 2.5%, 1.25%, and 0.25% of the final mixtures' total weight. Four MudPIT assays were acquired for each aliquot as described by Liu *et al*. [8]. Finally, we generated three sparse matrices to simulate three benchmarking scenarios; in the first scenario, the input vectors originating from the 25% protein spiking were labeled as belonging to the positive class and all the rest as to the negative class. On the second scenario, the 25% and the 2.5% input vectors were labeled as from the positive class and the rest as from the negative class. Finally, the third scenario labels the 25%, 2.5% and 1.25% input vectors as belonging to the positive class and the remaining 0.25% as belonging to the negative class.

Each sparse matrix was normalized according to the Z method and nSVM was applied to predict which and how many markers were spiked in the first matrix (scenario 1) using the linear SVM kernel and varying some parameters of the GA (Table 1). Almost all parameter combinations correctly top-ranked the spiked markers, and pinpointed how many markers were spiked in the lysate. The dataset used for testing originated from a 12 salt step MudPIT of a whole cell yeast lysate having more than 1800 identified proteins; this is far more complex than the average proteomic experiment. Therefore, more combinatorial possibilities were available, increasing computation time. Given these facts, nSVM is expected to perform faster in less complex studies (with fewer features). The island mode and mutation index play a key role in the GA; while apparently there are no great changes in execution time, runs using a mutation index of 2 with the island mode turned on yielded better results in our dataset. We then opted to use the island model and a mutation index of 2 to evaluate nSVM over the other two scenarios; the

Table 1: Evaluation of nSVM results on the spiking dataset using different parameters

Elitism	Mutation	Islands	No. Feat.	Avg. No. Subst.	Time	PHS2	ALB	CAH	ITRA
0	2	0	6	290	49 ± 9	1	3	5	4
0	2	250	4	380	50 ± 8	1	2	4	3
0	3	0	5	400	54 ± 6	2	1	3	5
0	3	250	4	415	51 ± 9	1	3	5	6
1	1	250	4	280	49 ± 7	4	1	2	3
1	2	0	6	360	56 ± 7	1	2	4	3
1	2	250	4	410	59 ± 3	1	3	4	2
1	3	0	5	416	54 ± 5	5	3	4	7
1	3	250	6	423	52 ± 7	1	6	3	2

nSVM was executed 20 times for each of the 8 specified conditions. PHS2, ALB, CAH, and ITRA stand for the spiked protein markers, and the numbers in the respective columns indicate the ranking. The Elitism column stands for how many individuals of the population were allowed to remain untouched for the following generation. The Islands column indicates how many seconds were required before a migration could even occur; a zero indicates that the island model was not applied. The No. Feat. column indicates what nSVM suggested as the minimum set of optimal features. The Avg. No. Subst. column indicates how many times the fittest individual was substituted. The Time column indicates the average time and standard deviation of 1 nSVM run. These results were obtained for scenario 1 as described in the Suggestion of when to apply nSVM section. Due to the stochastic nature of the method, results may vary.

method correctly predicted which markers were spiked in both cases (Table 2).

The Z normalization was chosen because one of the steps for estimating the VC dimension, according to the SVM *light* algorithm [21], is based on approximating the radius of the smallest hyphersphere that encompasses all input vectors by the norm of the largest support vector. After applying the Z normalization, the new mean for each PID becomes 0 and the data points become "evenly" distributed around the origin; this makes the VC dimension estimate more accurate. However, if the changes between samples are expected to be minimum, applying nSVM on "unnormalized" data can also be considered.

The widely adopted Student's t-test was then applied to check if we could rank the spiked markers as the topmost in the three scenarios after Z and Total Signal normalization. These results are listed in Table 3. By comparing the results from nSVM (Tables 1 and 2) against the t-test results (Table 3), it can be noted that the latter was unable to correctly rank the markers for some scenarios (all markers should have rank at most 4) and therefore did not reveal the optimal set. On the other hand, nSVM correctly ranked the spiked markers from all sparse matrices, which justifies the extra computation time it required. Recall that nSVM encompasses multiple executions (e.g., 20), and therefore more time to terminate (~1 h on an Intel Core 2 duo at 2.1 GHz). The limitations of the t-test seem to be that it relies on estimates of the mean and variance that do not necessarily reflect the true values when only a few samples are available [30]. nSVM's stochastic nature, combined with the various executions, makes it less prone to overfitting, but we note that it was unable to obtain the global optimum in any of the three matrices with only a single execution.

We recommend the t-test over nSVM for experiments where many changes are expected. Table 2 shows that even though the optimal result was not always achieved, very satisfactory results were obtained. Therefore, the t-test can offer a quick "bird's eye" view over changes throughout the entire experiment. On the other hand, nSVM works its way down to a minimum set of features, optimized for classification purposes, and therefore is probably not advisable for a holistic view.

Table 2: nSVM results in the spiking dataset (scenarios 2 and 3)

Scenario	PHS2	ALB	CAH	ITRA
2	4	1	3	2
3	2	1	3	4

PHS2, ALB, CAH, and ITRA stand for the spiked protein markers, and the numbers in the respective columns indicate the ranking according to nSVM.

Table 3: Student's t-test results for the spiking experiment

Normalization Method	PHS2	ALB	CAH	ITRA
	scenario 1			
Z	3	1	25	2
Total Signal	2	18	187	30
	scenario 2			
Z	10	1	2	3
Total Signal	1	2	4	3
	scenario 3			
Z	111	2	1	5
Total Signal	5	2	1	4

PHS2, ALB, CAH, and ITRA stand for the spiked protein markers, and the numbers in the respective columns indicate ranking according to Student's t-test applied to the sparse matrix normalized by Z or Total Signal normalization.

Differently than traditional GAs, nSVM offers a new strategy to estimate which proteins are differentially expressed. Our approach is a variation of the one used by Li *et al*. [31] to select differentially intensified mass spectral peaks from Surface Enhanced Laser Desorption Ionization – Time Of Flight proteomic profiles. Briefly, the authors repeatedly executed their GA rooted in k-nearest neighbor, a non-parametric pattern recognition method, to obtain relatively small subsets of discriminative mass spectral peaks between classes of specimens. Then peak appearance frequencies in the solutions were calculated and the authors showed that the most frequently selected peaks were the most discriminative. The efficiency of the algorithm was then proven on a validation set. The heuristics behind nSVM are far more computationally expensive than the one described by Li *et al*., so multiple executions (e.g., 1000) would invalidate its applicability. However, nSVM adopts a strategy that allows it to converge to very satisfactory solutions within only a few runs (e.g., 20).

Prior attempts at performing feature selection based solely on some function of the VC dimension [32] have been reported. However, our GA is based on the SRM principle that combines such a function with empirical error measures. Furthermore, we take advantage of a second theoretical error bound related to the number of support vectors to make nSVM converge faster (data not shown). We compared nSVM's performance with and without computing the empirical error measures; the former achieved better results on our dataset (data not shown).

Conclusion

The identification of trustworthy protein markers is not an easy task, since mass spectrometry based proteomics is still in development and spectral counting effectiveness can vary on the experimental setup, including mass spectrometry type and data-dependent analysis configuration. PatternLab implements several existing strategies and adds two new tools to the proteomic data analysis arsenal, each one having its own niche. Our results showed that even in simple scenarios, where the spiked concentrations can be considered relatively high, the data can still play tricks on well-founded feature selection methods. This is due to the dataset's high dimensionality, sparseness, and lack of a known *a priori* probability distribution. In even more realistic and complex scenarios, markers might be present in extremely low concentrations. Modification in the experimental designs to isolate sub-proteomes is a solution; however, these separations are many times not straightforward if protein content is to be disturbed only minimally. Therefore, even with all the advances in pattern recognition techniques, a set of *bona fide* markers requires experimental and computational validation in unseen samples to ensure the model is not a result of overfitting.

Availability and requirements

- **Project name:** PatternLab for proteomics

- **Project home page:** http://pcarvalho.com/patternlab

- **Operating system(s):** Windows XP or VISTA. PatternLab is expected soon to run under Linux and Macintosh, thanks to the Mono project [33].

- **Programming language:** C#

- **Other requirements:** .NET 3.5

- **License:** GNU

- **Any restrictions to use by non-academics:** license needed

Authors' contributions

PCC coded the software and wrote the first draft of the manuscript under the guidance of VCB and JRY. EIC and JSGF generated the MudPIT experimental, helped test the software, and suggested the inclusion of important features in it. All authors read and approved the final manuscript and contributed to the development of the project's website.

Acknowledgements

The authors acknowledge CNPq, CAPES, a FAPERJ BBP grant, the Rio de Janeiro proteomic network, and NIH P41RR011823 for financial support.

References

1. Jessani N, Niessen S, Wei BQ, Nicolau M, Humphrey M, Ji Y, Han W, Noh DY, Yates JR 3rd, Jefferey SS, Cravatt BF: **A streamlined platform for high-content functional proteomics of primary human specimens.** *Nat Methods* 2005, **2:**691-697.
2. Washburn MP, Wolters D, Yates JR III: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology.** *Nat Biotechnol* 2001, **19:**242-247.
3. Yates JR, Cociorva D, Liao L, Zabrouskov V: **Performance of a linear ion trap-Orbitrap hybrid for peptide analysis.** *Anal Chem* 2006, **78:**493-500.
4. Katajamaa M, Oresic M: **Processing methods for differential analysis of LC/MS profile data.** *BMC Bioinformatics* 2005, **6:**179.
5. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolster D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ: **A proteomic view of the Plasmodium falciparum life cycle.** *Nature* 2002, **419:**520-526.
6. Pang JX, Ginanni N, Dongre AR, Hefta SA, Opitek GJ: **Biomarker discovery in urine by proteomics.** *J Proteome Res* 2002, **1:**161-169.
7. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH: **Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards.** *Anal Chem* 2003, **75:**4818-4826.
8. Liu H, Sadygov RG, Yates JR III: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics.** *Anal Chem* 2004, **76:**4193-4201.
9. Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF: **Detecting differential and correlated protein expression in label-free shotgun proteomics.** *J Proteome Res* 2006, **5:**2909-2918.
10. Chen EI, Cociorva D, Norris JL, Yates JR III: **Optimization of mass spectrometry-compatible surfactants for shotgun proteomics.** *J Proteome Res* 2007, **6:**2529-2538.
11. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res* 1997, **7:**986-995.
12. Vapnik VN: *The Nature of Statistical Learning Theory* New York: Springer-Verlag; 1995.
13. Eng JK, McCormack AL, Yates JR III: **An approach to correlate tandem mass spectra data of peptides with amino acid sequences in a protein database.** *Anal Chem* 1995, **67(8):**1426-1436.
14. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20:**3551-3567.
15. Tabb DL, McDonald WH, Yates JR III: **DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics.** *J Proteome Res* 2002, **1:**21-26.
16. Cheadle C, Vawter MP, Freed WJ, Becker KG: **Analysis of microarray data using Z score transformation.** *J Mol Diagn* 2003, **5:**73-81.
17. Cleary JG, Teahan WJ: **Experiments on the zero frequency problem.** In *Proceedings of the Conference on Data Compression: 28–30 March 1995; Snowbird, UT* Edited by: Storer JA, Cohn M. Los Alamitos, CA: IEEE Computer Society Press; 1995:480.
18. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B* 1995, **57:**289-300.
19. Holland JH: *Adaptation in Natural and Artificial Systems* Ann Arbor, MI: University of Michigan Press; 1974.
20. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene selection for cancer classification using support vector machines.** *Mach Learn* 2002, **46:**389-422.
21. Joachims T: **Making large-scale SVM learning practical.** In *Advances in Kernel Methods: Support Vector Learning* Edited by: Schölkopf B, Burges C, Smola A. Cambridge, MA: MIT Press; 1999:169-184.
22. Carvalho PC, Carvalho MG, Degrave W, Lilla S, De NG, Fonseca R, Spector N, Musacchio J, Domont GB: **Differential protein expression patterns obtained by mass spectrometry can aid in the diagnosis of Hodgkin's disease.** *J Exp Ther Oncol* 2007, **6:**137-145.
23. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
24. Yang YH, Xiao Y, Segal MR: **Identifying differentially expressed genes from microarray experiments via statistic synthesis.** *Bioinformatics* 2005, **21:**1084-1093.
25. Yeung KY, Bumgarner RE, Raftery AE: **Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data.** *Bioinformatics* 2005, **21:**2394-2402.

26. Aubert J, Bar-Hen A, Daudin JJ, Robin S: **Determination of the differentially expressed genes in microarray experiments using local FDR.** *BMC Bioinformatics* 2004, **5:**125.
27. **Effects of perillyl alcohol on glioblastoma multiform cells (A172)** [http://pcarvalho.com/patternlab/downloads/poh]
28. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK: **Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.** *Bioinformatics* 2003, **19:**1945-1951.
29. Cover TM, Van Campenhout JM: **On the possible orderings in the measurement selection problem.** *Transactions on Systems, Man and Cybernetics* 1977, **7:**657-661.
30. Jafari P, Azuaje F: **An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors.** *BMC Med Inform Decis Mak* 2006, **6:**27.
31. Li L, Umbach DM, Terry P, Taylor JA: **Application of the GA/KNN method to SELDI proteomics data.** *Bioinformatics* 2004, **20:**1638-1640.
32. Fröhlich H, Chapelle O, Schölkopf B: **Feature selection for support vector machines using genetic algorithms.** *Int J Artif Intell Tools* 2004, **13:**791-800.
33. **Mono project** [http://www.mono-project.com]

ANEXO III

Online News

PatternLab for differential shotgun proteomics

Article Tools

Email a Colleague

Permalink

Recommend & Share

CiteULike

Delicious

Digg This

Facebook

Newsvine

This Toolbox item describes an algorithm, database, or computational method for proteomics analyses.

Laura Cassiday

Publication Date (Web): December 2, 2008

A major goal of proteomics is to identify differences in protein expression in various biological states. However, a one-size-fits-all approach to protein quantitation in LC/LC/MS/MS experiments is not always successful. Therefore, Paulo Carvalho and colleagues at the Federal University of Rio de Janeiro and Scripps Research Institute developed PatternLab for proteomics, a program that implements existing strategies and two new methods for protein quantitation by spectral counting.

PatternLab unifies a variety of strategies for feature selection and normalization, each of which has its own niche. One of the new data analysis tools, ACFold, is useful for experiments with fewer than three replicates from each state or for data obtained with multiple protocols (e.g., methods that use different MS-compatible detergents during proteolytic digestion). The second new tool for relative protein quantitation, a natural support vector machine, is ideal for the analysis of a minimum set of proteins (e.g., proteins that characterize a given pathology for use in an early detection kit). PatternLab is available for free at http://pcarvalho.com/patternlab. (*BMC Bioinf*. **2008**, DOI 10.1186/1471-2105-9-316)

1155 Sixteenth Street N.W., Washington, DC 20036

ACS Publications is a partner of: crossref COUNTER PORTICO

Technology Partner - Atypon Systems, Inc.

ANEXO IV

PubMed

U.S. National Library of Medicine
National Institutes of Health

ELSEVIER FULL-TEXT ARTICLE

Display Settings: Abstract

J Proteomics. 2010 Mar 10;73(5):1018-27. Epub 2010 Jan 18.

Dynamic proteomic overview of glioblastoma cells (A172) exposed to perillyl alcohol.

Fischer Jde S, Liao L, Carvalho PC, Barbosa VC, Domont GB, Carvalho Mda G, Yates JR 3rd.

Laboratory for Protein Chemistry, Chemistry Institute, and the Rio de Janeiro Proteomic Network, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.

Perillyl alcohol (POH) is a naturally occurring terpene and a promising chemotherapeutic agent for glioblastoma multiform; yet, little is known about its molecular effects. Here we present results of a semi-quantitative proteomic analysis of A172 cells exposed to POH for different time-periods (1', 10', 30', 60', 4h, and 24h). The analysis identified more than 4000 proteins; which were clustered using PatternLab for proteomics and then linked to Ras signaling, tissue homeostasis, induction of apoptosis, metallopeptidase activity, and ubiquitin-protein ligase activity. Our results make available one of the most complete protein repositories for the A172. Moreover, we detected the phosphorylation of GSK3beta (Glycogen synthase kinase) and the inhibition of ERK's (extracellular signal regulated kinase) phosphorylation after 10', which suggests a new mechanism of POH's activation for apoptosis. Copyright (c) 2010 Elsevier B.V. All rights reserved.

PMID: 20083244 [PubMed - in process]

PMCID: PMC2834810 [Available on 2011/3/10]

Publication Types, Grant Support

LinkOut - more resources

You are here: NCBI > Literature > PubMed

Write to the Help Desk

ANEXO V

Analyzing Shotgun Proteomic Data with PatternLab for Proteomics

Paulo C. Carvalho,[1,2] John R. Yates III,[2] and Valmir C. Barbosa[1]

[1] Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
[2] The Scripps Research Institute, La Jolla, California

ABSTRACT

PatternLab for proteomics is a one-stop shop computational environment for analyzing shotgun proteomic data. Its modules provide means to pinpoint proteins/peptides that are differentially expressed and those that are unique to a state. It can also cluster the ones that share similar expression profiles in time-course experiments, as well as help in interpreting results according to Gene Ontology. PatternLab is user-friendly, simple, and provides a graphical user interface. *Curr. Protoc. Bioinform.* 30:13.13.1-13.13.14. © 2010 by John Wiley & Sons, Inc.

Keywords: shotgun proteomics • label-free proteomic analysis • label-based proteomic analysis.

INTRODUCTION

Shotgun proteomics is a powerful approach for analyzing complex peptide mixtures. The overall strategy comprises the digestion of proteins followed by peptide separation, fragmentation, and protein identification (Washburn et al., 2002). Tandem mass spectra are acquired to enable protein identification, which is commonly achieved by comparing experimental with theoretically generated spectra and pinpointing the most likely match via search engines such as SEQUEST (see *UNIT 13.3*; Eng et al., 1994), ProLuCID (Xu et al., 2006), or Mascot (Perkins et al., 1999). The identifications are then filtered according to quality scores and a false-discovery rate is estimated. SEQUEST or ProLuCID followed by DTASelect (see *UNIT 13.4*; Cociorva et al., 2007) are the search engines and filtering program, respectively, used in the protocols listed herein.

Proteins are quantitated according to label-free or label-based (e.g., SILAC; Ong et al., 2002) strategies. For example, spectral counting is a label-free method that works with the numbers of identified spectra matched to a protein (Liu et al., 2004); it is simple and has been shown to be successful on controlled experiments with spiked proteins (Carvalho et al., 2008b). Label-based strategies yield higher confidence for relative quantitation; nevertheless, they are more expensive and laborious. The latter are performed by comparing a peptide to an internal, chemically identical standard enriched with a heavy stable isotope; the ratios informing relative abundance are then obtained computationally with a program such as Census (see *UNIT 13.12*; Park et al., 2008).

In general, protein identification and quantitation constitute the tip of the iceberg for analyzing shotgun proteomic data. Questions such as “Which proteins are differentially expressed?”, “Which are unique to a state?”, “Which share similar expression profiles in a time-course experiment?”, and even more specific questions, such as, “Which proteins originate from the mitochondria?” are all very common. PatternLab for proteomics (Carvalho et al., 2008a) is a one-stop shop for answering these types of question. PatternLab provides tools for analyzing shotgun proteomic data quantitated by spectral counting (output from DTASelect) or by the label-based (output from Census). Its modules provide

Current Protocols in Bioinformatics 13.13.1-13.13.14, June 2010
Published online June 2010 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/0471250953.bi1313s30

ANEXO VI

PubMed

U.S. National Library of Medicine
National Institutes of Health

FULL TEXT ONLINE WILEY InterScience FREE Author Manuscript in PubMed Central

Display Settings: Abstract

Plant J. 2010 Jan;61(2):290-9. Epub 2009 Oct 26.

PYR/PYL/RCAR family members are major in-vivo ABI1 protein phosphatase 2C-interacting proteins in Arabidopsis.

Nishimura N, Sarkeshik A, Nito K, Park SY, Wang A, Carvalho PC, Lee S, Caddell DF, Cutler SR, Chory J, Yates JR, Schroeder JI.

Division of Biological Sciences, Cell and Developmental Biology Section and Center for Molecular Genetics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0116, USA.

Abscisic acid (ABA) mediates resistance to abiotic stress and controls developmental processes in plants. The group-A PP2Cs, of which ABI1 is the prototypical member, are protein phosphatases that play critical roles as negative regulators very early in ABA signal transduction. Because redundancy is thought to limit the genetic dissection of early ABA signalling, to identify redundant and early ABA signalling proteins, we pursued a proteomics approach. We generated YFP-tagged ABI1 Arabidopsis expression lines and identified in vivo ABI1-interacting proteins by mass-spectrometric analyses of ABI1 complexes. Known ABA signalling components were isolated including SnRK2 protein kinases. We confirm previous studies in yeast and now show that ABI1 interacts with the ABA-signalling kinases OST1, SnRK2.2 and SnRK2.3 in plants. Interestingly, the most robust in planta ABI1-interacting proteins in all LC-MS/MS experiments were nine of the 14 PYR/PYL/RCAR proteins, which were recently reported as ABA-binding signal transduction proteins, providing evidence for in vivo PYR/PYL/RCAR interactions with ABI1 in Arabidopsis. ABI1-PYR1 interaction was stimulated within 5 min of ABA treatment in Arabidopsis. Interestingly, in contrast, PYR1 and SnRK2.3 co-immunoprecipitated equally well in the presence and absence of ABA. To investigate the biological relevance of the PYR/PYLs, we analysed pyr1/pyl1/pyl2/pyl4 quadruple mutant plants and found strong insensitivities in ABA-induced stomatal closure and ABA-inhibition of stomatal opening. These findings demonstrate that ABI1 can interact with several PYR/PYL/RCAR family members in Arabidopsis, that PYR1-ABI1 interaction is rapidly stimulated by ABA in Arabidopsis and indicate new SnRK2 kinase-PYR/PYL/RCAR interactions in an emerging model for PYR/PYL/RCAR-mediated ABA signalling.

PMID: 19874541 [PubMed - in process]

PMCID: PMC2807913

Publication Types, Grant Support

LinkOut - more resources

You are here: NCBI > Literature > PubMed

Write to the Help Desk

ANEXO VII

Proteome Science

BioMed Central

Methodology

Open Access

GO Explorer: A gene-ontology tool to aid in the interpretation of shotgun proteomics data

Paulo C Carvalho†[1,2], Juliana SG Fischer†[2,3], Emily I Chen*[2,6], Gilberto B Domont[3], Maria GC Carvalho[4], Wim M Degrave[5], John R Yates III[2] and Valmir C Barbosa[1]

Address: [1]Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Brazil, [2]Department of Chemical Physiology, The Scripps Research Institute, La Jolla, USA, [3]Chemistry Institute, Federal University of Rio de Janeiro, and Rio de Janeiro Proteomics Network, Rio de Janeiro, Brazil, [4]Carlos Chagas Filho Biophysics Institute, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, [5]Oswaldo Cruz Institute, Laboratory for Functional Genomics and Bioinformatics, Rio de Janeiro, Brazil and [6]Department of Pharmacological Sciences, Stony Brook University, Stony Brook, NY, USA

Email: Paulo C Carvalho - paulo@buscario.com.br; Juliana SG Fischer - juli_f@iq.ufrj.br; Emily I Chen* - emily@pharm.stonybrook.edu; Gilberto B Domont - gilberto@iq.ufrj.br; Maria GC Carvalho - mgccosta@biof.ufrj.br; Wim M Degrave - wdegrave@fiocruz.br; John R Yates - jyates@scripps.edu; Valmir C Barbosa - valmir@cos.ufrj.br

* Corresponding author   †Equal contributors

Published: 24 February 2009

Proteome Science 2009, **7**:6   doi:10.1186/1477-5956-7-6

Received: 26 November 2008
Accepted: 24 February 2009

This article is available from: http://www.proteomesci.com/content/7/1/6

Abstract

Background: Spectral counting is a shotgun proteomics approach comprising the identification and relative quantitation of thousands of proteins in complex mixtures. However, this strategy generates bewildering amounts of data whose biological interpretation is a challenge.

Results: Here we present a new algorithm, termed GO Explorer (GOEx), that leverages the gene ontology (GO) to aid in the interpretation of proteomic data. GOEx stands out because it combines data from protein fold changes with GO over-representation statistics to help draw conclusions. Moreover, it is tightly integrated within the PatternLab for Proteomics project and, thus, lies within a complete computational environment that provides parsers and pattern recognition tools designed for spectral counting. GOEx offers three independent methods to query data: an interactive directed acyclic graph, a specialist mode where key words can be searched, and an automatic search. Its usefulness is demonstrated by applying it to help interpret the effects of perillyl alcohol, a natural chemotherapeutic agent, on glioblastoma multiform cell lines (A172). We used a new multi-surfactant shotgun proteomic strategy and identified more than 2600 proteins; GOEx pinpointed key sets of differentially expressed proteins related to cell cycle, alcohol catabolism, the Ras pathway, apoptosis, and stress response, to name a few.

Conclusion: GOEx facilitates organism-specific studies by leveraging GO and providing a rich graphical user interface. It is a simple to use tool, specialized for biologists who wish to analyze spectral counting data from shotgun proteomics. GOEx is available at http://pcarvalho.com/patternlab.

Background

Shotgun proteomics is a strategy capable of identifying thousands of proteins in complex mixtures. Its methodology comprises the pre-digestion of proteins followed by peptide separation, fragmentation in a mass spectrometer, and database search [1,2]. Multi-dimensional Protein Identification Technology (MudPIT) is a shotgun proteomics technique capable of identifying thousands of proteins in proteolytically digested complex mixtures [2,3]. MudPIT separates peptides according to two independent physicochemical properties using two-dimensional liquid chromatography (LC/LC) online with the ion source of a mass spectrometer. This separation relies on columns of strong cation exchange (SCX) and reversed phase (RP) material, back to back, inside fused silica capillaries. The chromatography proceeds in cycles, each of which consists of increasing salt concentration to "bump" peptides off the SCX followed by a hydrophobic gradient to progressively elute peptides from the RP into the ion source. This process identifies mixture components by tandem mass spectrometry (MS/MS). Relative protein quantitation can be obtained through tandem mass spectral features (e.g., peptide hits, protein sequence coverage, spectral counts) [1,3-5]. For example, Liu *et al.* demonstrated that the number of tandem mass spectra obtained for each protein, or "spectral count", linearly correlates with its abundance in a mixture by two orders of magnitude [6]. Currently, spectral counting is a widely adopted approach to characterize different states of biological systems according to protein expression differences.

Acquiring a holistic understanding over a large set of proteins is not a simple task, but first insights can be obtained by searching the Gene Ontology (GO) [7] annotations for over-represented terms. GO is a standard for functional annotation and consists of structured and controlled vocabularies to classify terms into the following root categories (namespaces): molecular function, biological processes, and cellular components. Its structure follows that of a directed acyclic graph (DAG); each term is a more specific child of one or more parents (i.e., directed edges point in the direction of increasing specificity). In this way, a convention named *true path rule* states that whenever a gene is annotated with a term, it is also implicitly annotated with all (less specific) ancestors of that term.

Currently, there are several GO-based tools; some examples are: DAVID [8], GOMiner [9], and GoFish [10]. We refer the reader to http://www.geneontology.org/GO.tools.shtml for a more comprehensive listing. Even though such tools are frequently used to analyze microarray data, the ones specific for proteomics amount to very few [11]. Moreover, most existing GO-based tools for proteomics overlook expression fold changes and, as far as we know, are not specialized in directly handling data from differential proteomic spectral counting experiments. One exception with relation to the use of fold changes is GESA (Gene Enrichment Analysis) [12], which ranks genes according to expression quantitation data and then correlates them to search for enriched GO terms. However, limiting the search to enriched terms can hide very subtle results elucidated by individual proteins. In this respect, we note that GOEx provides several exploratory methods that are not bound to finding terms that are necessarily enriched but could be related even to one single protein.

In this work we present a new GO-based tool, named GO Explorer (GOEx), which is optimized to work with spectral counting data from shotgun proteomics. This is achieved, in part, because GOEx is natively integrated into the PatternLab for Proteomics project [13] so it leverages existing parsers, data normalization, and feature selection algorithms designed to work with spectral counts. GOEx allows one to explore data using several new approaches as described in the Implementation section.

We demonstrate GOEx by using proteomic data acquired from human glioblastoma multiform (GBM) cell lines (A172) both before and after applying perillyl alcohol (POH) to their medium. Briefly, POH is a naturally occurring monoterpene found in lavender, cherries, and mint, and is a promising chemotherapeutic agent. In human cancer cells, POH has shown cytostatic and cytotoxic effects [14-16], inducing apoptosis on lung [17], leukemia [18], prostate [19], and breast [20] cancer cell lines. POH is also under evaluation in several clinical trials, including an ongoing phase I comprising GBM patients treated by intranasal delivery that has shown promising results [21].

Experimental: preparation of the A172-POH dataset

Materials

Invitrosol™ and RapiGest™ SF acid-labile surfactant were purchased from Invitrogen (Carlsbad, CA) and Waters Corp. (Milford, MA), respectively. PPS Slient surfactant was provided by Dr. Norris from Protein Discovery, Inc. (Knoxville, TN). The proteases endoproteinase Lys-C and trypsin (modified, sequencing grade) were obtained from Roche. Human malignant glioma cells (A172) were obtained from the American Type Culture Collection. POH and other laboratory reagents were purchased from Sigma-Aldrich (St. Louis, MO), unless noted otherwise.

Cell culture and POH treatment

The A172 cells were grown as monolayers in 25 cm^2 tissue culture flasks in Dulbecco's modified Eagle medium supplemented with 0.2 mM non-essential amino acids, 10% fetal calf serum, penicillin (60 μg/mL), streptomycin (100 μg/mL), and amphotericin B (fungizone, 2.5 mg/mL). For sub-cultivations, confluent monolayers were gently washed with phosphate-buffered saline (PBS 1×) pH 7.2, and after short trypsinization the cells were suspended in culture medium. Three subcultures were treated with 1.8 mM POH (Sigma-Aldrich, 96%) during 1.5 h and three other subcultures received no POH treatment; the cellular morphology analyzed by an optical phase-contrast microscope (Zeiss Axioplan, Thornwood, NY) and the cells were photographed. The medium from all cultures was discarded and the cells were rinsed twice with PBS (1×). The cells were detached from the flask by exposing them during 2 min in a solution of 0.25% trypsin-EDTA (1×). Then the cells were re-suspended in the medium and a pellet was obtained by centrifugation during 10 min at 500 RCF. This procedure was performed three times. Proteins were extracted from the cell pellets using the total protein extraction kit from Biochain (Hayward, CA) according to manufacturer's instructions.

Protein solubilization with MS-compatible detergents and trypsin digestion

Each protein pellet was re-suspended, independently, with one of the following MS-compatible detergents: 5 μL of Invitrosol (5× stock), RapiGest SF (1% stock), or 10 μL of PPS (1% stock). We recall that these detergents are called MS-compatible because they do not interfere with the mass spectral acquisition, increase proteolytic efficiency, and peptide and protein identifications in complex protein mixtures analyzed by shotgun proteomics [22]. The concentration of each detergent used in this study was determined based on the maximum recommended concentration suggested by the manufacturers. Then the proteins were incubated at 60°C for 5 min and completed with solvent (PPS reconstituted in the same buffer, RapiGest reconstituted in 50 mM ammonium bicarbonate, Invitrosol is already sold in solution) to a 50 μL final volume. All samples were sonicated for 2 h in a water bath and digested with trypsin (1:50) for 16 h at 37°C.

Post-digestion

Following digestion, all reactions were acidified with 90% (v/v) formic acid (2% final) to stop the proteolysis. Samples with RapiGest SF and PPS were acidified and incubated at 37°C for additional 4 h to facilitate the hydrolysis of the detergents. Then samples were centrifuged for 30 min at 14,000 rpm to remove insoluble material. The soluble peptide mixtures were collected, dried by a Speed Vac, reconstituted in 10 μL of buffer A (95% H_2O (v/v), 5% acetonitrile (v/v), and 0.1% formic acid (v/v)), and analyzed by MudPIT[1].

Protein identification by MudPIT

Approximately 70 μg of the digested peptide mixture were loaded onto a biphasic (strong cation exchange/reversed phase) capillary column and washed with a buffer containing 5% acetonitrile, 0.1% formic acid diluted in HPLC grade water. The two-dimensional liquid chromatography separation and tandem mass spectrometry conditions were as described by Washburn *et al.* [1]. The flow rate at the tip of the biphasic column was 300 nL/min when the mobile phase composition was 95% H_2O, 5% acetonitrile, and 0.1% formic acid. The ion trap mass spectrometer, Finnigan LCQ Deca XP (Thermo Finnigan, San Jose, CA), was set to the data-dependent acquisition mode with dynamic exclusion turned on. One MS survey scan was followed by four MS/MS scans and 12 salt steps were performed. Mass spectrometer scan functions and HPLC solvent gradients were controlled by the Xcalibur data system (Thermo Finnigan, San Jose, CA).

Tandem mass spectra were extracted from the raw files, and a binary classifier, previously trained on a manually validated dataset, was used to remove the low-quality MS/MS spectra [23]. The remaining spectra were searched against the *Homo sapiens* protein plus common contaminant proteins; all sequences were downloaded as FASTA-formatted from the EBI-IPI protein database (database version 3.23, released on November 2, 2006) [24]. To calculate confidence levels and false-positive rates, a decoy database that contained the reverse sequences of the original dataset appended to the target database was used [25], and the best matching sequences from the combined database were indicated by SEQUEST [26]. The searches were done on a cluster of Intel Xeon 80 processors running the Linux operating system. The peptide mass search tolerance was set to 3 Da. No differential modifications were considered. For the aqueous digestion, the mass of the amino acid cysteine was statically modified by +57 Da due to the carboxyamidomethylation of the sample. No enzymatic cleavage conditions were imposed on the database search, so the search space included all candidate peptides whose theoretical mass fell within the 3 Da mass tolerance window, regardless of their tryptic status.

The validity of peptide/spectrum matches was assessed in DTASelect 2 [27] according to the SEQUEST cross-correlation score (XCorr) and the SEQUEST normalized difference in cross-correlation score (DeltaCN). The search results were grouped by charge state (+1, +2, and +3) and tryptic status (fully tryptic, half-tryptic, and non-tryptic), resulting in 9 distinct subgroups. In each of the subgroups, the distribution of XCorr and DeltaCN values for the direct and decoy database hits was obtained, and the two subsets were separated by quadratic discriminant analysis. Outlier points in the two distributions (for example, matches with very low XCorr but very high DeltaCN) were discarded. Full separation of the direct and decoy subsets is not generally possible; therefore, the discriminant score was set such that a false-discovery rate of 5% was determined based on the number of accepted decoy database peptides. This procedure was independently performed on each data subset, resulting in a false-positive rate independent of tryptic status or charge state. In addition, a minimum sequence length of 7 amino-acid residues was required, and each protein on the list was supported by at least two peptide identifications unless specified otherwise. These additional requirements, especially the latter, resulted in the elimination of most decoy database and false-positive hits, as these tended to be overwhelmingly present as proteins identified by single peptide matches. After this last filtering step, the estimated false-discovery rate was reduced to below 1%.

Selecting differentially expressed proteins with PatternLab's ACFold

ACFold is part of the PatternLab for Proteomics project [13] and considers information from protein fold changes, the AC test [28], and a false-discovery rate (FDR) estimator [29] to pinpoint differentially expressed proteins. We recall that the AC test can be used to calculate the conditional probability of finding a spectral count of x_2 in biological state 2 given that a spectral count of x_1 was observed in biological state 1. The ACFold method was chosen because it is designed to search for differential protein patterns in shotgun proteomic data and can be applied even if the assays are not technical replicates, as in our multi-surfactant shotgun proteomic approach [13,22].

ACFold is effective because drawing conclusions using only a fold change cutoff can shadow information from low-level protein changes that might be important. To account for such, ACFold relies on the AC test to fish out proteins that, despite not having achieved a theoretical optimal fold-cutoff, do nevertheless exhibit a difference in spectral counts between states that is statistically significant. Such proteins are put in evidence to be re-considered in the final analysis or for further experimental validation.

We refer to Figure 1 to illustrate the output of PatternLab's ACFold graphical user interface and also further details of an ACFold analysis. We also remark that, additionally, PatternLab incorporates the TFold method, in which the t-test substitutes for the AC test, for use when 3 or more replicate readings for each state are available.

In this work, PatternLab's parser was used to convert the DTASelect files from all MudPIT assays into the unified PatternLab format before loading them to the ACFold tool. An FDR q-value of 0.1 and an AC test p-value of 0.05 were specified. The Row Sigma normalization [13] was chosen for computing the fold changes. The fold change cutoff of 2.5 was empirically specified so as to maximize the number of proteins that satisfy both the FDR and the AC test criteria. We note that higher fold change cutoffs reduce the number of verified hypotheses, usually increasing (decreasing) the number of proteins approved by the FDR (AC test). Finally, a report listing the proteins that satisfied all criteria (ACFoldReport) was exported to text format. This report is also the input to the GOEx analysis. We refer the reader to Figure 1 to illustrate PatternLab's ACFold graphical user interface output of the identified proteins' distribution.

Implementation

GO Explorer

GOEx was coded using C# 3.5 and carried a graphical user interface for improved user experience. GOEx requires the downloading of two files: the latest GO ontology (OBO v.1.2 format), freely available at http://www.geneontology.org/GO.downloads.ontology.shtml, and the GOA (gene ontology annotation) association file containing the non-redundant, species-specific annotation, freely available at http://www.ebi.ac.uk/GOA/goaHelp.html. The latter is necessary to convert the IPI's (international protein indexes), obtained during protein identification, into the GO terms. In this work, we used the gene_ontology_edit.obo (Feb. 08, 2008) and the gene_association.goa_human (Feb. 03, 2008) files. From then on, GOEx parses both files and performs various pre-computations (e.g., mapping all terms descending from a specific term) and associations to speed up the user's experience when analyzing data. All information is then compacted into a binary representation, in a process known as serialization, and saved to disk for quick retrieval during a future use.

Finally, the GOEx panel is unlocked and the GO root terms are listed in the interactive directed acyclic graph (iDAG) interface. The user can then load a report of the differentially expressed proteins (e.g., ACFoldReport) to be analyzed in any of the GOEx study modes: iDAG-driven, specialist-driven, and automatically driven. For convenience, henceforth we refer to the proteins reported in the ACFoldReport as "reported proteins".

Calculating the over-representation p*-value*

First the accession number listed in the "reported proteins" file are converted into their equivalent GO terms. This conversion entails a mapping that can occur at different levels of the GO hierarchy (not only at the leaves) and sometimes a protein can be mapped onto more than one GO term. While the conversion takes place, tags are maintained for each term indicating which proteins were mapped onto it.

The over-representation p-value of term termed as S relative to the namespace of source (least specific term) G is computed as follows. Let g denote the total number of GO terms in the namespace of source G and let s - 1 be the

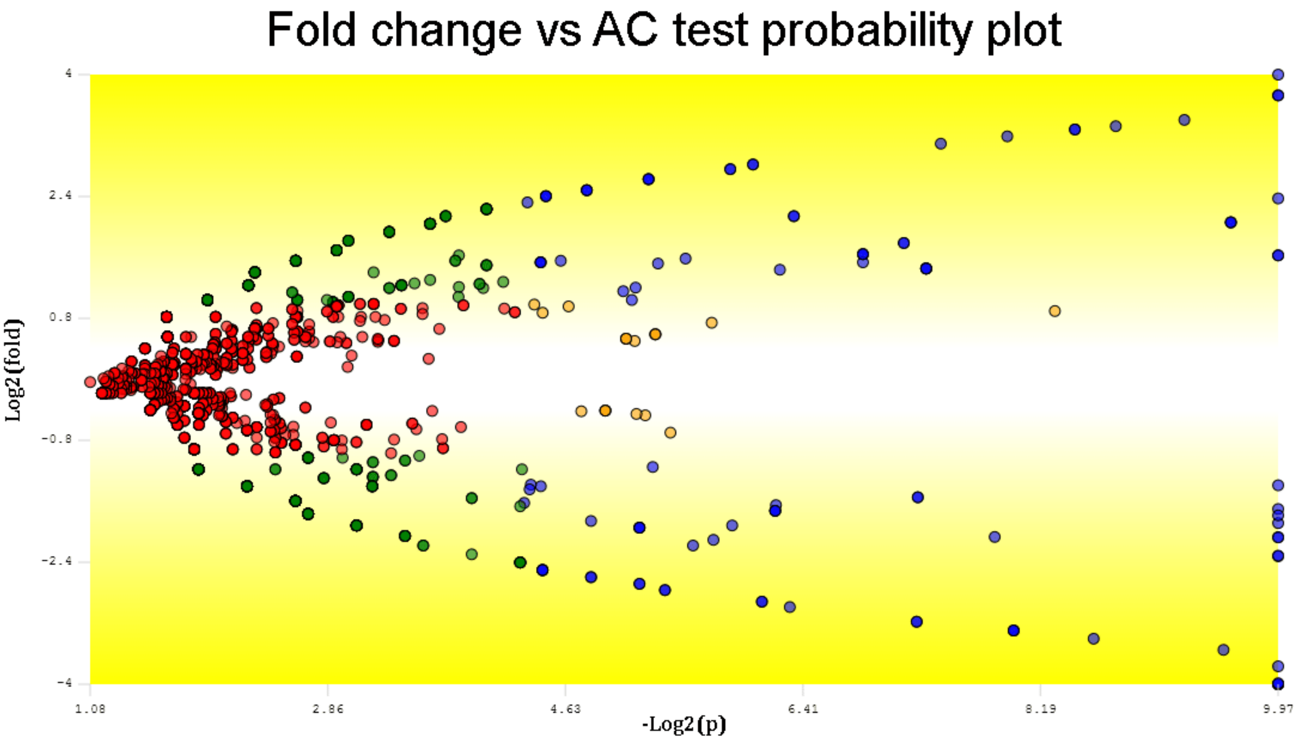


Figure 1

Fold change versus AC test probability plot. This plot was obtained using PatternLab's ACFold algorithm and displays the results obtained with the multi-surfactant shotgun proteomic approach when comparing the A172 cell lines before and after the treatment with perillyl alcohol. Each protein (represented as a dot) was mapped according to its $\log_2$(fold change) on the ordinate (y) axis and $-\log_2(1-(\text{AC test } p\text{-value}))$ on the abscissa (x) axis. A total of 104 proteins (blue dots) were selected as differentially expressed because they satisfied both the AC test and the FDR *q*-value specified cutoffs. 23 proteins (orange dots) did not meet the fold change cutoff but were indicated as statistically differentially expressed, therefore deserving further analysis. 267 proteins (green dots) met the fold change cutoff, but the AC test indicated that this happened by chance. 2293 proteins (red dots) were pinpointed as not differentially expressed between classes because they failed both the AC test and the fold change cutoffs. The number of dots does not match the number of identified proteins due to the many overlaps.

number of GO terms that descend from S–thus s includes S itself and its descent. The overrepresentation p-value of S must be computed so as to reflect the distinct proteins that were mapped onto the s terms. Counting the number of such proteins from the tags maintained during the mapping process is not enough because, in principle, the result may amount to more than s. Letting $c(S)$ be this number of distinct proteins, the count we actually use is then $k = \min\{s, c(S)\}$. The probability of observing these k distinct proteins for a randomly selected S can now be estimated by the hypergeometric distribution: if X is the corresponding random variable, then

$$P(X = k) = \frac{\binom{t}{k}\binom{g-t}{s-k}}{\binom{g}{s}},$$

where $t = \min\{g, c(G)\}$, following the same reasoning that led to the definition of k. This given, we express the overrepresentation p-value of term S as the probability of observing k or more distinct proteins mapped onto the s terms, that is,

$$P(X \geq k) = 1 - \sum_{i=0}^{k-1} \frac{\binom{t}{i}\binom{g-t}{s-i}}{\binom{g}{s}}.$$

Clearly, the lower this p-value the greater the probability mass that lies strictly below k.

Data Analysis

a) The GOEx iDAG-driven mode

This strategy is designed to help guide one's biological questions by leveraging the GO through the iDAG coupled with graphing tools. By clicking on an iDAG term, its child terms appear listed below it; for each child, its over-representation p-value (described above) and the sum of the protein fold changes reported for it are computed. Terms having no relation to the reported proteins are automatically deleted to keep the biological questions on track. A "distribution pie chart" (Figure 2A) and a "fold change versus over-representation plot" (Figure 2B) of the displayed iDAG leaf terms are presented. A report table discriminating all calculations and the reported proteins related to each term is also made available. All this information can aid in choosing which term to explore next, if any, thus helping drive one's biological questions. In general, terms having low p-values and/or high-magnitude fold changes are good candidates, but there are important exceptions. For example, while exploring for putative molecular functions of our reported proteins, we noted that the "molecular transducer activity" GO term presented a significant fold change (*fold* = 12) but was not statistically over-represented (p = 0.97). Even so, by further expanding it and examining its child terms, GOEx revealed the "G-protein coupled receptor activity" term (*fold* = -8, p = 0.89) to be associated with our reported proteins, which is only as expected according to previous work related to the effects of POH on tumor cells [21]. This example illustrates that it is possible to draw the important conclusions from fold change data only. GO tools, however, tend to overlook this, being generally limited to taking into account over-representation p-values exclusively.

b) The GOEx specialist-driven mode

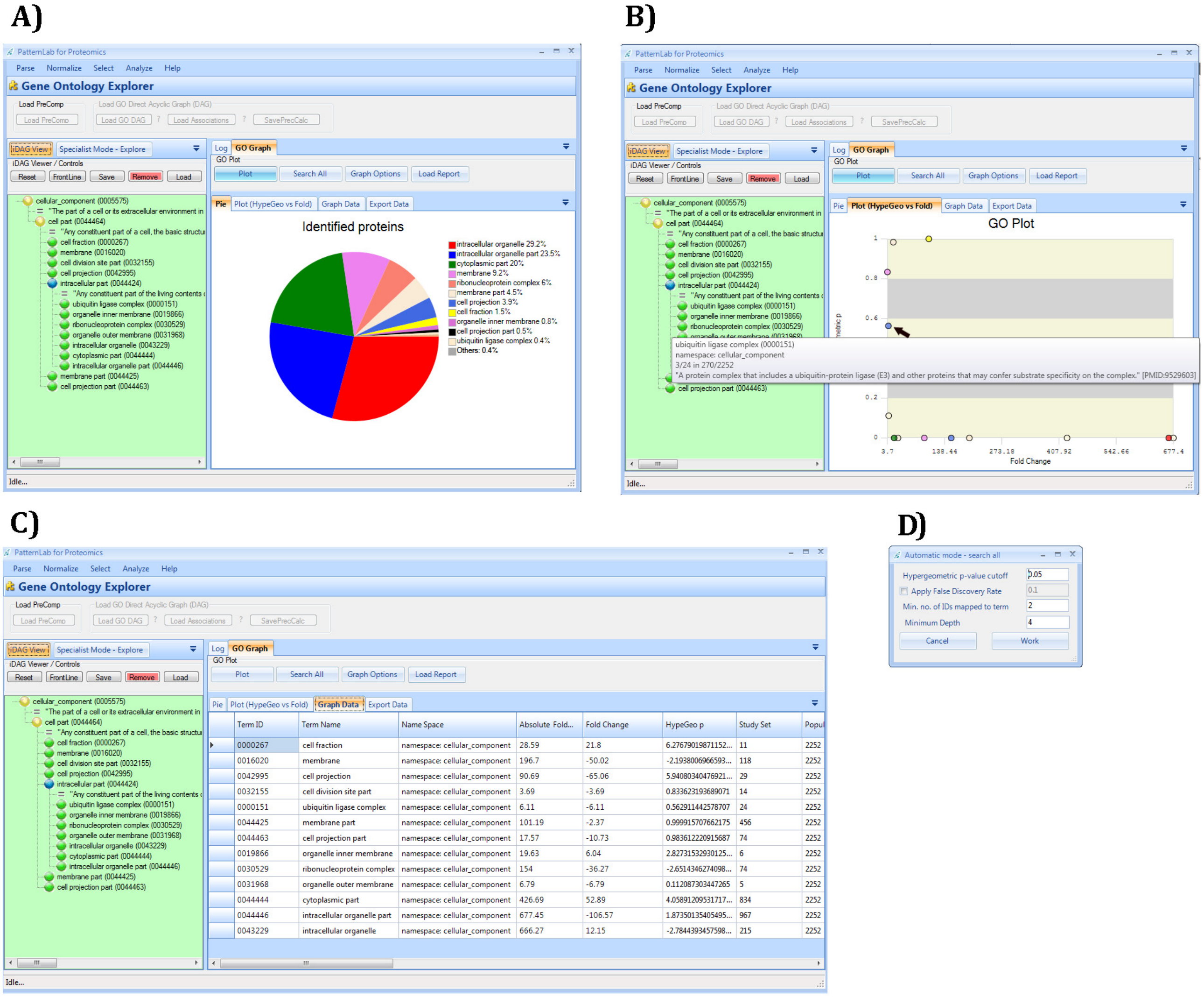
This mode allows an expert to pose questions and retrieve answers in the light of the GO and the reported proteins. For example, it is known that the Ras signaling pathway has a key role in the pathogenesis of GBM by acting as a primary switch that mediates external signals to numerous intracellular signaling pathways [30]. It is also known that POH affects the levels of Ras-related proteins and Ras isoprenylation, thereby altering cellular physiology [31]. Entering the key word "Ras" to the search facility of the GOEx specialist-driven mode produced, in a log file, a list of all GO terms containing the key word in their names or descriptions. Terms related to the reported proteins (either through fold change or over-representation) were analyzed, plotted, and added to the report table. The result pointed to the "Ras protein signal transduction" term as being related to our dataset despite not quite qualifying as statistically over-represented (p = 0.06 against a p-value cutoff of 0.05). This example indicates that, even though a term's over-representation may not be indisputably significant (and thus the term might not be detected during an automatic search, as in most GO tools), that term may nevertheless embody the correct answer. In the case at hand, the literature gives plenty of supporting evidence to corroborate the hypothesis of alterations in the Ras pathway. This is further addressed in the Results and discussion section.

c) The GOEx automatic mode

The automatic mode (search all) performs an extensive analysis by searching for relations between the reported proteins and each and every GO term. This method requires the user to specify the desired minimum number of proteins related to a GO term, a minimum GO depth, and an over-representation p-value and optionally a false-discovery rate [29]) cutoff. We define GO depth as the shortest path from a term to its root. From then on, GOEx will evaluate all GO terms. The ones bearing relation to the reported proteins will be listed in the report table (described in section d) and plotted in the "distribution pie chart" and "fold change versus over-representation plot". This mode is optimized for multi-core processors and relies on concurrent computation to speed up its task.

d) The GOEx report table

All GOEx query methods provide the already mentioned complementary report table that can be dynamically sorted according to convenience. The table headers include: GO ID, Term Name, Namespace, Absolute Fold Change, Fold Change, HypeGeo P, Study Set, Population, Identified in Study Set, Identified in Population, Proteins IPI's and Folds, GO depth, and Description. GO ID and Term Name specifies the unique GO identifier and its name as given in the GO. Namespace points to which GO namespace the selected term belongs to (molecular function, cellular component, or biological process). Absolute Fold Change is the sum of the absolute values of the fold changes of all proteins mapped onto a given GO term. Similarly, Fold Change is the sum all their fold change values. Current gene ontology tools usually do not report fold change information. HypeGeo P is an abbreviation for the term's over-representation p-value. Study Set refers to all the terms that descend from a given term. Population stands for all the terms contained within the specified term's namespace. Identified Proteins indicates how


A)
PatternLab for Proteomics
Parse Normalize Select Analyze Help
Gene Ontology Explorer
Load PreComp
Load GO Direct Acyclic Graph (DAG)
Load PreComp
Load GO DAG
Load Associations
SavePrecCalc
iDAG View
Specialist Mode - Explore
iDAG Viewer / Controls
Reset
FrontLine
Save
Remove
Load
cellular_component (0005575)
"The part of a cell or its extracellular environment in
cell part (0044464)
"Any constituent part of a cell, the basic structu
cell fraction (0000267)
membrane (0016020)
cell division site part (0032155)
cell projection (0042995)
intracellular part (0044424)
"Any constituent part of the living contents c
ubiquitin ligase complex (0000151)
organelle inner membrane (0019866)
ribonucleoprotein complex (0030529)
organelle outer membrane (0031968)
intracellular organelle (0043229)
cytoplasmic part (0044444)
intracellular organelle part (0044446)
membrane part (0044425)
cell projection part (0044463)
Log
GO Graph
GO Plot
Plot
Search All
Graph Options
Load Report
Pie
Plot (HypeGeo vs Fold)
Graph Data
Export Data
Identified proteins
🟥 intracellular organelle 29.2%
🟦 intracellular organelle part 23.5%
🟩 cytoplasmic part 20%
membrane 9.2%
ribonucleoprotein complex 6%
membrane part 4.5%
cell projection 3.9%
🟨 cell fraction 1.5%
organelle inner membrane 0.8%
⬛ cell projection part 0.5%
ubiquitin ligase complex 0.4%
Others: 0.4%
Idle...
B)
GO Plot
1
0.8
0.6
0.2
0
3.7
138.44
273.18
407.92
542.66
677.4
Fold Change
Idle...
C)
D)

ubiquitin ligase complex (0000151)
namespace: cellular_component
3/24 in 270/2252
"A protein complex that includes a ubiquitin-protein ligase (E3) and other proteins that may confer substrate specificity on the complex." [PMID:9529603]

Term ID	Term Name	Name Space	Absolute Fold...	Fold Change	HypeGeo p	Study Set	Popul
0000267	cell fraction	namespace: cellular_component	28.59	21.8	6.27679019871152...	11	2252
0016020	membrane	namespace: cellular_component	196.7	-50.02	-2.1938006966593...	118	2252
0042995	cell projection	namespace: cellular_component	90.69	-65.06	5.94080340476921...	29	2252
0032155	cell division site part	namespace: cellular_component	3.69	-3.69	0.833623193689071	14	2252
0000151	ubiquitin ligase complex	namespace: cellular_component	6.11	-6.11	0.562911442578707	24	2252
0044425	membrane part	namespace: cellular_component	101.19	-2.37	0.999915707662175	456	2252
0044463	cell projection part	namespace: cellular_component	17.57	-10.73	0.983612220915687	74	2252
0019866	organelle inner membrane	namespace: cellular_component	19.63	6.04	2.82731532930125...	6	2252
0030529	ribonucleoprotein complex	namespace: cellular_component	154	-36.27	-2.6514346274098...	74	2252
0031968	organelle outer membrane	namespace: cellular_component	6.79	-6.79	0.112087303447265	5	2252
0044444	cytoplasmic part	namespace: cellular_component	426.69	52.89	4.05891209531717...	834	2252
0044446	intracellular organelle part	namespace: cellular_component	677.45	-106.57	1.87350135405495...	967	2252
0043229	intracellular organelle	namespace: cellular_component	666.27	12.15	-2.7844393457598...	215	2252

Automatic mode - search all

Hypergeometric p-value cutoff: 0.05
☐ Apply False Discovery Rate: 0.1
Min. no. of IDs mapped to term: 2
Minimum Depth: 4

Cancel Work

Figure 2
The GOEx graphical user interface. A) A pie chart showing the distribution of the identified proteins as mapped onto selected cellular component GO terms is displayed on the right. The level of specificity was chosen according to the iDAG in the left panel. B) The GO terms related to the iDAG terms specified on the left are plotted according to the overrepresentation *p*-value and absolute fold change calculated for them from the identified proteins. The mouse is currently hovering over one term and its GO description is provided in a balloon. A detailed report table can be accessed by clicking on the Graph Data tab. C) Detailed information on the displayed results can be accessed by clicking on the Graph data tab. The table can be dynamically sorted by clicking on the column of interest. A detailed description of each column is addressed in The GOEx report table section. D) The automatic search pop-up window appears when one clicks on the Search all button in the main interface. The user can then select several stringency values to search for statistically overrepresented terms.

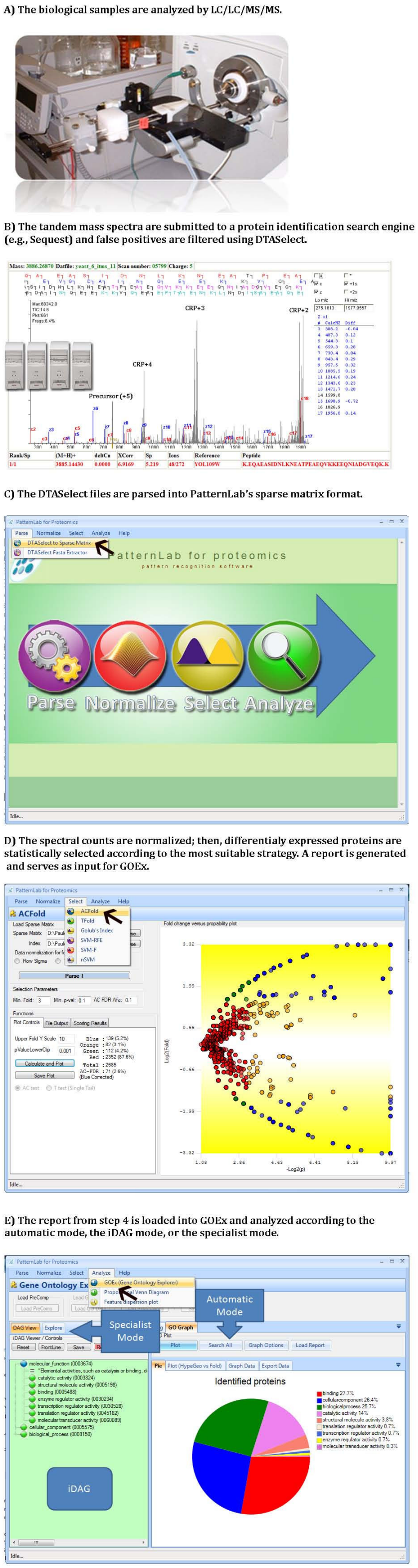
many of the proteins discriminated in the ACFoldReport were mapped onto the specified term. Proteins IPI's and Folds discriminates all the proteins, and respective fold changes, mapped onto the selected term. Finally, Description refers to the term's GO description.

Results and discussion

We refer the reader to Figure 3 for an illustration of the main steps that led to the results we now present.

Protein identification by the multi-surfactant shotgun proteomic approach

Protein solubility varies in different buffers and in the presence of different types of detergents. Therefore, pro-



A) The biological samples are analyzed by LC/LC/MS/MS.

B) The tandem mass spectra are submitted to a protein identification search engine (e.g., Sequest) and false positives are filtered using DTASelect.

C) The DTASelect files are parsed into PatternLab's sparse matrix format.

D) The spectral counts are normalized; then, differentialy expressed proteins are statistically selected according to the most suitable strategy. A report is generated and serves as input for GOEx.

E) The report from step 4 is loaded into GOEx and analyzed according to the automatic mode, the iDAG mode, or the specialist mode.

Figure 3
Workflow. Key steps in the workflow, ranging from the mass spectral acquisition to the final GOEx analysis.

tein solubilization by different MS-compatible detergents can provide complementary data [22]. In this way, our multi-surfactant proteomic approach can potentially cover a larger portion of the proteome than the traditional technical replicate approach, and improve the GO analysis [22]. Our proteomic methodology identified a total of 2687 proteins during all six MudPIT runs and PatternLab's ACFold selected 104 of them as differentially expressed. An additional 23 proteins that did not satisfy our fold cutoff but had a very low AC test *p*-value (the ACFold orange group) were independently evaluated and included in our list.

As far as we know, our A172-POH dataset is the largest one concerning GBM A172 cells. Such repository, together with the DTASelect files and the reported differentially expressed proteins, is available for download at the PatternLab for Proteomics project website and can be a valuable source to test future GO approaches. Taken together, the proteins identified in the present study can also provide important fundamental information about the cellular response to POH treatment.

The GOEx specialist mode results

The "Ras protein signal transduction" term was linked to two proteins: transforming protein RhoA (IPI00478231) and Rho-related GTP-binding protein RhoB (IPI00000041). RhoA is involved in regulating the signal transduction pathway between the plasma membrane receptors for the assembly of focal adhesions and actin stress fibers. Yan and collaborators have reported RhoA's expression to positively correlate with the degree of malignancy in astrocytomas and that its expression is increased in various neoplasias. The authors also suggest important implications of RhoA in both the clinical prognosis and the biology of these neoplasms, and even suggest using it as a prognostic biomarker [32]. Our results showed a down-regulation of ~3× for RhoA after the POH treatment, showing POH to be effective as a chemotherapeutic agent.

RhoB was also down-regulated (~4×) after the POH treatment. RhoB is linked with endothelial cell survival during angiogenesis and has been hypothesized to have a role in TNFalpha-induced angiogenesis through the regulation of Akt activation, being therefore important for tissue repair during acute inflammatory responses [33]. Thus, the fact that POH is an angiogenesis inhibitor is in agreement with our results [34]. Moreover, the authors also report that inhibiting the farnesylation of RhoB is a strategy for treatment. Indeed, one of the key effects of POH is to inhibit the farnesylation of Ras proteins, preventing them from docking in the plasma membrane and initiating signal transduction [21].

The GOEx automatic search result

The GOEx automatic mode can provide complementary results to the specialist when compared to the iDAG-driven mode, as exemplified in the Implementation section. We performed an automatic search on our dataset using an FDR of 0.05 and eliminating terms that had each only one protein assigned to it. The results pointed mostly to terms related to cell cycle, alcohol catabolism, the Ras pathway, apoptosis, and stress response. Examples of terms belonging to the molecular function namespace and selected as overrepresented include, but are not limited to: purine nucleotide binding, hydrolase activity–acting on acid anhydrides–in phosphorus-containing anhydrides, structural molecule activity, and cytoskeletal protein binding. Similarly, terms belonging to the cellular component namespace include, but are not limited to: membrane-bound vesicle, actin filament bundle, cytoskeletal part, and cytosolic part. Finally, terms from the biological process namespace include, but are not limited to: microtubule-based process, actin filament-based process, regulation of apoptosis, alcohol metabolic process, and GTP metabolic process. Indeed, apoptosis, changes in morphology, and most of the terms listed were only expected, according to previous work [16,20,21]; microscopy images of the cells can be found in Figure 4.

The GOEx methodology

There are several methods to compute an over-representation *p*-value; examples are: the hypergeometric [35], binomial, χ^2 (chi-square), and Fisher's exact [36] tests; their differences have been reported not to be dramatic for the GO overrepresentation problem [37]. Most GO-based tools are limited to what is equivalent to the GOEx automatic search in terms of limiting the search to finding statistically over-represented terms. To speed up their analyses, they usually do not offer over-representation calculation using the hypergeometric distribution and/or use GO-slim, a reduced version of GO. However, analyses according to the latter are restricted to the higher GO levels, which contrast sharply with our approach, which takes into account all levels and every term. This limitation could lead to missing differences that are detectable only at more refined levels. With the advent of faster microprocessors, the time to complete a full GO search has dramatically decreased, so what was once considered an issue to worry about has been downshifted. Nevertheless, GOEx also takes advantage of the new multi-core chips to perform concurrent computing to accelerate the automatic search.

Even though variations on how to find over-represented terms can be proposed, there is no reference standard on how to properly measure the gains. So comparisons between methods are bound, to some extent, to be disputed [38]. GOEx stands out among other methods because it lies within a complete workflow to analyze shotgun proteomic experiments that rely on spectral counting. Most importantly, its reports combine information from fold changes with statistics. As we exemplified, these two types of information are complementary, yet most existing GO tools do not take this fact into account. In any given biological phenomenon, different genes are regulated to different extents. The data providing information about differential protein expression can be useful in assigning different weights to the corresponding biological processes involved and aid in inferring which biological process is more relevant [37]. Certainly, the greatest limitation of GOEx, and of all existing GO-based tools as well, is that GO, the IPI database, and the mappings, all of which serve as foundations for such tools, are not complete, which evidently affects the results they yield. Such

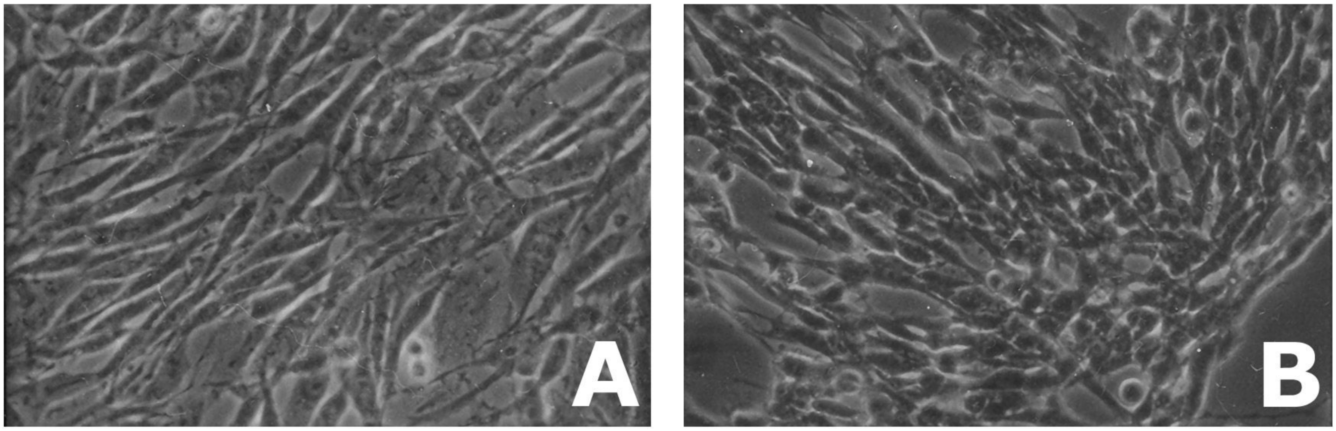


Figure 4
Microscopy images of the A172 cells. These microscopy images (200×) show the A172 cell line before (A) and after treatment with POH during 1.5 h. The cellular morphology changes and the cells become rounder after the POH treatment.

limitation is inevitable but tends to become less important as these databases are expanded.

In all, GOEx provides several strategies to explore how the proteins of interest are distributed among GO terms. Differently than the automatically driven methods of previous software, GOEx embodies flexible exploratory tools. For example, as terms are expanded in the iDAG, child terms onto which any identified protein is mapped are kept even if not statistically enriched. This retains terms that could contain a single protein and yet be crucial for drawing conclusions. Thus, GOEx's iDAG or specialist mode can determine both whether GO categories are statistically over-represented and whether there are significant changes for individual proteins.

It seems to be a consensus that web-based tools are more liable because the researcher can be assured to be using the software's latest version: maintaining a stand-alone installation represents one more chore to the user. However, the GOEx installation has been designed to be straightforward; in fact, it can be done with one single click of the mouse. If the application needs upgrades or detects any missing components, they are automatically downloaded. Nevertheless, if a major change has been deployed but the user is unsatisfied, a rollback (restore) can be done in one single step, differently than the web-based case, in which one is forced to use the available version. In this way, GOEx provides benefits in a locally installed distribution, besides not forcing the user to share sensitive data with an unknown and remote server. In conclusion, GOEx facilitates organism-specific searches using GO through a rich graphical user interface. It is a useful, friendly, and simple to use tool, specialized for biologists who wish to analyze spectral counting data from shotgun proteomics.

Availability and requirements

GOEx is available for download at http://pcarvalho.com/patternlab and is free for academic use. It was programmed in C# and requires .NET 3.5 framework (can be automatically installed) and a windows (VISTA or XP) personal computer.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PCC coded the software and wrote the first draft of the manuscript under the guidance of VCB and JRY. EIC and JSGF generated the MudPIT experimental data, prepared the POH-A172 cells, helped test the software, and suggested the inclusion of important features. WMD discussed several aspects of the software. GBD and MGCC participated during all phases as JSGF's doctoral advisers. All authors read and approved the final manuscript.

Acknowledgements

This work was funded by CNPq, CAPES, a FAPERJ BBP grant, PAPES, PDTIS, the Ary Frauzino Foundation, NIAID sub-contract grant UCSD/MCB0237059, and partially supported by NIH P41RR011823, NIH 5R01 MH067880, the genesis molecular biology laboratory, and the Fiocruz-INCA collaboration.

References

1. Washburn MP, Wolters D, Yates JR III: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology.** *Nat Biotechnol* 2001, **19:**242-247.
2. Yates JR, Cociorva D, Liao L, Zabrouskov V: **Performance of a linear ion trap-Orbitrap hybrid for peptide analysis.** *Anal Chem* 2006, **78:**493-500.
3. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ: **A proteomic view of the Plasmodium falciparum life cycle.** *Nature* 2002, **419:**520-526.
4. Pang JX, Ginanni N, Dongre AR, Hefta SA, Opitek GJ: **Biomarker discovery in urine by proteomics.** *J Proteome Res* 2002, **1:**161-169.
5. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH: **Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards.** *Anal Chem* 2003, **75:**4818-4826.
6. Liu H, Sadygov RG, Yates JR III: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics.** *Anal Chem* 2004, **76:**4193-4201.
7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Tarver LI, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
8. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4:**3.
9. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Jane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barret JC, Weinstein JN: **GoMiner: a resource for biological interpretation of genomic and proteomic data.** *Genome Biol* 2003, **4:**R28.
10. Berriz GF, White JV, King OD, Roth FP: **GoFish finds genes with combinations of Gene Ontology attributes.** *Bioinformatics* 2003, **19:**788-789.
11. Feng W, Wang G, Zeeberg BR, Guo K, Fojo AT, Kane DW, Reinhold WC, Lababidi S, Weinstein JN, Wang MD: **Development of gene ontology tool for biological interpretation of genomic and proteomic data.** *Proceedings of the AMIA Annual Symposium: 8–12 November 2003; Washington, DC* 2003:839.
12. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP: **GSEA-P: a desktop application for Gene Set Enrichment Analysis.** *Bioinformatics* 2007, **23:**3251-3253.
13. Carvalho PC, Fischer JSG, Chen EI, Yates JR III, Barbosa VC: **PatternLab for proteomics: a tool for differential shotgun proteomics.** *BMC Bioinformatics* 2008, **9:**316.
14. Clark SS, Zhong L, Filiault D, Perman S, Ren Z, Gould M, Yang X: **Anti-leukemia effect of perillyl alcohol in Bcr/Abl-transformed cells indirectly inhibits signaling through Mek in a Ras- and Raf-independent fashion.** *Clin Cancer Res* 2003, **9:**4494-4504.
15. Clark SS, Perman SM, Sahin MB, Jenkins GJ, Elegbede JA: **Antileukemia activity of perillyl alcohol (POH): uncoupling apoptosis from G0/G1 arrest suggests that the primary effect of POH on Bcr/Abl-transformed cells is to induce growth arrest.** *Leukemia* 2002, **16:**213-222.
16. Burke YD, Ayoubi AS, Werner SR, McFarland BC, Heilman DK, Ruggeri BA, Crowell PL: **Effects of the isoprenoids perillyl alcohol**

and farnesol on apoptosis biomarkers in pancreatic cancer chemoprevention. *Anticancer Res* 2002, **22:**3127-3134.

17. Yeruva L, Pierre KJ, Elegbede A, Wang RC, Carper SW: **Perillyl alcohol and perillic acid induced cell cycle arrest and apoptosis in non small cell lung cancer cells.** *Cancer Lett* 2007, **257:**216-226.
18. Clark SS: **Perillyl alcohol induces c-Myc-dependent apoptosis in Bcr/Abl-transformed leukemia cells.** *Oncology* 2006, **70:**13-18.
19. Chung BH, Lee HY, Lee JS, Young CY: **Perillyl alcohol inhibits the expression and function of the androgen receptor in human prostate cancer cells.** *Cancer Lett* 2006, **236:**222-228.
20. Yuri T, Danbara N, Tsujita-Kyutoku M, Kiyozuka Y, Senzaki H, Shikata N, Kanzaki H, Tsubura A: **Perillyl alcohol inhibits human breast cancer cell growth in vitro and in vivo.** *Breast Cancer Res Treat* 2004, **84:**251-260.
21. Da Fonseca CO, Landeiro JA, Clark SS, Quirico-Santos T, da Costa Carvalho MG, Gattass CR: **Recent advances in the molecular genetics of malignant gliomas disclose targets for antitumor agent perillyl alcohol.** *Surg Neurol* 2006, **65(Suppl 1):**S1.
22. Chen EI, Cociorva D, Norris JL, Yates JR III: **Optimization of mass spectrometry-compatible surfactants for shotgun proteomics.** *J Proteome Res* 2007, **6:**2529-2538.
23. Bern M, Goldberg D, McDonald WH, Yates JR III: **Automatic quality assessment of peptide tandem mass spectra.** *Bioinformatics* 2004, **20(Suppl 1):**i49-i54.
24. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R: **The International Protein Index: an integrated database for proteomics experiments.** *Proteomics* 2004, **4:**1985-1988.
25. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP: **Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome.** *J Proteome Res* 2003, **2:**43-50.
26. Yates JR III: **Database searching using mass spectrometry data.** *Electrophoresis* 1998, **19:**893-900.
27. Tabb DL, McDonald WH, Yates JR III: **DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics.** *J Proteome Res* 2002, **1:**21-26.
28. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res* 1997, **7:**986-995.
29. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B* 1995, **57:**289-300.
30. Feldkamp MM, Lau N, Guha A: **Signal transduction pathways and their relevance in human astrocytomas.** *J Neurooncol* 1997, **35:**223-248.
31. Crowell PL, Ren Z, Lin S, Vedejs E, Gould MN: **Structure-activity relationships among monoterpene inhibitors of protein isoprenylation and cell proliferation.** *Biochem Pharmacol* 1994, **47:**1405-1415.
32. Yan B, Chour HH, Peh BK, Lim C, Salto-Tellez M: **RhoA protein expression correlates positively with degree of malignancy in astrocytomas.** *Neurosci Lett* 2006, **407:**124-126.
33. Fernandez-Borja M: **RhoB regulates TNFalpha-induced Akt activation and angiogenesis.** *Vascular Pharmacology* **45:**e44.
34. Loutrari H, Hatziapostolou M, Skouridou V, Papadimitriou E, Roussos C, Kolisis FN, *et al.*: **Perillyl alcohol is an angiogenesis inhibitor.** *J Pharmacol Exp Ther* 2004, **311:**568-575.
35. Cho RJ, Huang M, Campbell MJ, Dong H, Steinmetz L, Sapinoso L, Hampton G, Elledge SJ, Davis RW, Lockhart DJ: **Transcriptional regulation and function during the human cell cycle.** *Nat Genet* 2001, **27:**48-54.
36. Man MZ, Wang X, Wang Y: **POWER_SAGE: comparing statistical tests for SAGE experiments.** *Bioinformatics* 2000, **16:**953-959.
37. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, **21:**3587-3595.
38. Grossmann S, Bauer S, Robinson PN, Vingron M: **Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis.** *Bioinformatics* 2007, **23:**3024-3031.

ANEXO VIII

PubMed

U.S. National Library of Medicine
National Institutes of Health

Display Settings: Abstract

Anal Chem. 2009 Feb 9. [Epub ahead of print]

Charge Prediction Machine: Tool for Inferring Precursor Charge States of Electron Transfer Dissociation Tandem Mass Spectra.

Carvalho PC, Cociorva D, Wong CC, Carvalho MD, Barbosa VC, Yates JR.

Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Brazil, Biological Mass Spectrometry Laboratory, The Scripps Research Institute, La Jolla, California, and Laboratory for Control of Gene Expression, Biophysics Institute, Federal University of Rio de Janeiro, Brazil.

Electron transfer dissociation (ETD) can dissociate highly charged ions. Efficient analysis of ions dissociated with ETD requires accurate determination of charge states for calculation of molecular weight. We created an algorithm to assign the charge state of ions often used for ETD. The program, Charge Prediction Machine (CPM), uses Bayesian decision theory to account for different charge reduction processes encountered in ETD and can also handle multiplex spectra. CPM correctly assigned charge states to 98% of the 13 097 MS2 spectra from a combined data set of four experiments. In a comparison between CPM and a competing program, Charger (ThermoFisher), CPM produced half the mistakes.

PMID: 19203245 [PubMed - as supplied by publisher]

LinkOut - more resources

ANEXO IX

BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 20 2009, pages 2734–2736
doi:10.1093/bioinformatics/btp489

Sequence analysis

YADA: a tool for taking the most out of high-resolution spectra

Paulo C. Carvalho[1,2,*], Tao Xu[2], Xuemei Han[2], Daniel Cociorva[2], Valmir C. Barbosa[1] and John R.Yates, III[2]

[1]Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Caixa Postal 68511, 21941-972 Rio de Janeiro - RJ, Brazil and [2]Department of Chemical Physiology, The Scripps Research Institute, N Torrey Pines 10550, La Jolla, CA 92037, USA

Received on June 26, 2009; revised on August 9, 2009; accepted on August 11, 2009

Advance Access publication August 14, 2009

Associate Editor: John Quackenbush

ABSTRACT

Summary: YADA can deisotope and decharge high-resolution mass spectra from large peptide molecules, link the precursor monoisotopic peak information to the corresponding tandem mass spectrum, and account for different co-fragmenting ion species (multiplexed spectra). We describe how YADA enables a pipeline consisting of ProLuCID and DTASelect for analyzing large-scale middle-down proteomics data.

Availability: http://fields.scripps.edu/yada

Contact: paulo@pcarvalho.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

High-resolution, high-mass-accuracy (<10 p.p.m. error) mass spectrometry allows highly charged peptide isotopic peaks to be distinguished from one another, thus enabling the calculation of their precise charge states and monoisotopic masses. This is crucial for confident protein identification, especially when dealing with large molecules, such as the ones usually obtained from transmembrane protein digests. Other key advantages when analyzing large molecules include an increased identification coverage and the possibility of assessing relationships among multiple modifications in the same molecule (e.g. histone; Chi *et al.*, 2007). These motivations have given rise to a new proteomics platform, termed middle-down (MD) proteomics, which focuses on large molecules usually obtained through proteases other than trypsin or by modifying the digestion protocols (e.g. short-time trypsin digests; Forbes *et al.*, 2001).

Traditional protein identification search engines [e.g. SEQUEST (Eng *et al.*, 1994) and Mascot (Perkins *et al.*, 1999)] cannot take full advantage of high-resolution mass spectra, especially those of large molecules, mainly because of three reasons: (i) The specific peak in an isotopic envelope being selected for fragmentation can have a significant difference in mass as compared with its monoisotope; this can lead the search engine astray. (ii) Fragmenting highly charged precursors results in highly charged daughter ions. If the charge states of the latter are unknown, all charge state hypotheses should be considered for protein identification, resulting in a combinatorial explosion that can burden the search engine severely and decrease chances of a successful identification. (iii) Approximately 10% of all tandem spectra from complex mixtures are composed of different ion species that are co-fragmenting; only one of these ion species is searched (Carvalho *et al.*, 2009).

In order to overcome these limitations, we introduce YADA, a tool that can deconvolute (i.e. deisotope and decharge) high-resolution spectral data of peptide ions having charges up to +18 or masses up to 20 kDa. Because deconvolution entails the assignment of a charge state and the recalculation of the m/z's as if the charge were +1 (decharging), the combinatorial explosion problem mentioned above is automatically eliminated. YADA can also update the MS2 to reflect the fragmented precursor monoisotopic mass by locating its corresponding isotopic envelope in the MS1 and replacing its m/z with the monoisotopic m/z. Accordingly, multiple precursors can be considered when peaks from different isotopic envelopes are found within the precursor isolation bounds; this enables the accounting for multiplexed spectra. As far as we know, no other freely available deisotoping and decharging tools offer both features.

We describe a freely available pipeline to address large-scale MD studies consisting of YADA; ProLuCID, a protein identification search engine that is ready to efficiently handle deconvoluted spectra (Xu *et al.*, 2006); and DTASelect, which controls and estimates the false discovery rates (Cociorva *et al.*, 2007). Again, we are aware of no freely available solution for such in the context of MD. Our pipeline is evaluated on a short-time trypsin digest of a yeast lysate. The results are compared with those obtained by replacing YADA with Xtract (Thermo, San Jose, CA, USA), a software for decharging based on the THRASH algorithm (Horn *et al.*, 2000), which is also present in a commercial solution for MD analysis [Thermo's ProsightPC v2.0 (Boyne *et al.*, 2009)].

2 ALGORITHM

2.1 Peak filtering

YADA filters noise peaks and peaks that can be eliminated without compromising isotopic envelope recognition. Filtering increases speed (by up to 40%), reduces RAM requirements and improves charge assignment (by ∼5%; data not shown). It is accomplished in two steps, as follows.

The first step eliminates peaks that fall below an intensity threshold. The threshold can by default be a user-specified hard

*To whom correspondence should be addressed.

cutoff, or else be automatically determined for each spectrum (recommended for top-down datasets). The latter is accomplished by treating each mass spectrum as a two-component probability mixture model. The premises are that the noise peaks have intensities that follow a normal distribution; and that peptide-derived peaks have considerably higher intensities which, though to a lesser extent than the noise peaks, can also be assumed to follow a normal distribution. Given these, the well-established expectation–maximization (EM) algorithm is employed to maximize the likelihood of the observed intensity histogram under the assumed bimodal normal distribution. The two EM seeds (starting points) are chosen by sorting the intensities of all peaks and choosing the ones that, globally, rank at 10% and 90%. Given the two normal distributions provided by EM, the threshold is equal to that of the optimal Bayesian classifier.

The second step discards peaks that do not contribute to charge determination. The premise in this case is that the intensities of peaks derived from a peptide ion isotope will monotonically increase until a local maximum is achieved, at which point they will monotonically decrease. The algorithm proceeds as follows. Mass spectral peaks are sorted by increasing m/z and an empty result array is created. Then, for every peak, if its intensity is greater than that of the previous peak and the two differ in m/z by less than a given p.p.m. tolerance (e.g. 30 p.p.m.), then the current peak replaces the latest peak included in the result array; otherwise the peak is simply included in the result array. Then the spectral peaks are sorted by decreasing m/z and the process is repeated.

2.2 Detecting and decharging an isotopic envelope

All peaks are candidate seeds for an isotopic envelope. Briefly, for a given peak, first a peak-finding algorithm is employed to integrate peak intensities by stepping a distance of 1.0024 divided by a charge state hypothesis, for several charge state hypotheses (e.g. +1 through +21). Second, only the charge state hypothesis to have found the greatest number of sequential peaks and having the greatest accumulated intensity is retained. The observed profile is normalized to 1 and its dot product with a normalized averagine theoretical profile for the estimated mass (obtained using a kernel regressor) is computed. If the dot product is above a given threshold, the envelope is stored. The candidate envelope is discarded if verified to be part of an existing envelope of the same charge or one that would produce an overlapping envelope (e.g. +4 and +2). The algorithm is not subtractive and is capable of identifying overlapping envelopes.

2.3 Decharging, clustering and accounting for multiplexed spectra

It is common to observe the same peptide ion species with different charge states in the same spectrum. After deconvolution, these peptides yield peaks that are very close (e.g. <0.2 Da apart for a given resolution) to one another. YADA automatically coalesces them by averaging their masses and summing their intensities.

Also, in complex samples, it is common to have more than one ion species in the same isolation window to be fragmented when generating a tandem mass spectrum. YADA uses isotopic envelope information from the preceding MS1 to assign a monoisotopic precursor mass to the MS2 spectrum and to consider multiple ion species within the isolation window bounds (multiplexed spectra).

Table 1. Results

	Xtract	Y	Y with Corr
Number of identified spectra	898	996	1071
Fraction of identified spectra with non-monoisotopic precursor assigned (%)	74	75	1

Y stands for YADA, Corr for monoisotopic correction and multiplexing.

3 RESULTS

A 30 min trypsin digest of a yeast lysate was analyzed with a 2 h LC-MS run, acquiring one high-resolution MS1 (60 000 resolution at 400 m/z) followed by three high-resolution ETD-MS2 scans on the Orbitrap (Makarov, 2000) in data-dependent mode (10 071 spectra). The spectra were extracted using RawExtract (McDonald *et al*., 2004) and processed by YADA (with and without monoisotopic correction and multiplexing), ProLuCID (protein identification) and DTASelect. The latter ensures a peptide false discovery rate <1% against a decoy database. An example of a multiplexed spectrum solved by YADA (Supplementary Fig. S1) and further details regarding the search parameters and false discovery rate estimation are presented in the Supplementary Material. For evaluation purposes, an in-house script was created to use Xtract to deconvolute the data and replace YADA in our pipeline. The results are presented in Table 1. YADA turned out to be ~600% faster than the commercial software during deconvolution (YADA: 6′ 31″; Xtract: 42′ 24″; both on a 1 GHz Athlon with 1 GB RAM).

4 FINAL CONSIDERATIONS

While previous tools (Chen and Yap, 2008; Horn *et al*., 2000) are devoted solely to deconvolution, YADA's hallmark is its ability to maximize the results of large-scale experiments by quickly deconvoluting highly charged MD MS2 spectra and accounting for multiple precursors (multiplexed spectra).

We also note that, although it has been shown that assigning monoisotopic precursor masses to MS2's increases protein identification confidence (Mayampurath *et al*., 2008), MD poses a new challenge. This is so because large molecules (>12 kDa) often have the monoisotopic peak intensity below the detection sensitivity. Previous strategies have relied only on detected peaks, but YADA can predict a large molecule's undetected monoisotopic peak by considering its three most intense envelope peaks and estimating the monoisotopic mass according to an averagine model.

We have also described a freely available pipeline for analyzing high-resolution tandem mass spectra of large peptide molecules. The pipeline can be used for datasets containing high-resolution MS1 and MS2 spectra, or only a high-resolution MS1. In the case of the later, the MS2 cannot be deconvoluted; however, identification results can still be improved by assigning monoisotopic masses. The key steps are listed in Figure 1.

YADA is coded in C# and requires a PC with Windows XP SP2 or later and .NET 3.5. It installs under the directory PatternLab for proteomics (Carvalho *et al*., 2008). The windows version can be downloaded at our web site (http://fields.scripps.edu). A command-line version (requires MONO; http://mono-project.com), executable on Windows or Linux, is available upon request.

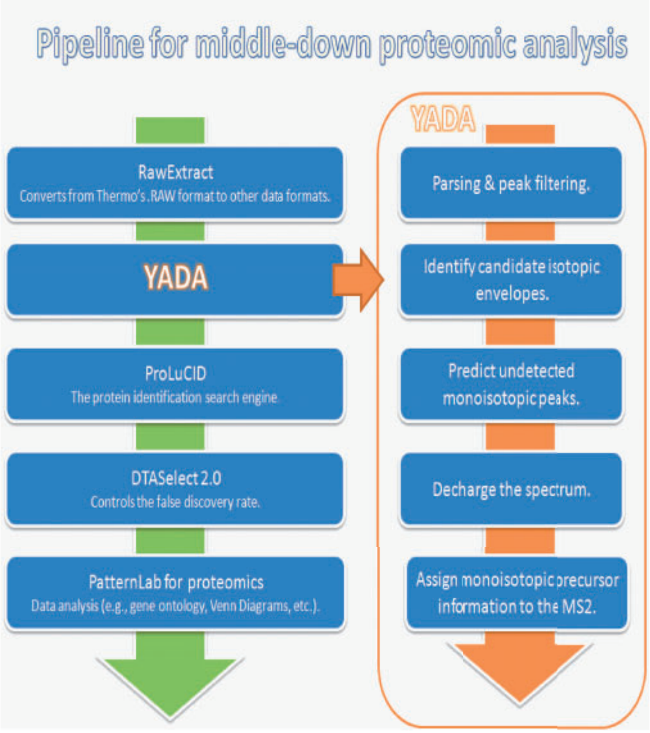

Pipeline for middle-down proteomic analysis
RawExtract
Converts from Thermo's .RAW format to other data formats.
YADA
ProLuCID
The protein identification search engine.
DTASelect 2.0
Controls the false discovery rate.
PatternLab for proteomics
Data analysis (e.g., gene ontology, Venn Diagrams, etc.).
YADA
Parsing & peak filtering.
Identify candidate isotopic envelopes.
Predict undetected monoisotopic peaks.
Decharge the spectrum.
Assign monoisotopic precursor information to the MS2.

Fig. 1. The key steps of the proposed pipeline for middle-down proteomic analysis.

YADA's current version is not recommended for the analysis of isotopically labeled datasets, since in these cases the isotopic distribution patterns may differ from the theoretically predicted.

Funding: This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (a FAPERJ BBP grant), and the National Institutes of Health [P41RR011823, ROI MH06880].

Conflict of Interest: none declared.

REFERENCES

Boyne,M.T. *et al*. (2009) Tandem mass spectrometry with ultrahigh mass accuracy clarifies peptide identification by database retrieval. *J. Proteome Res.*, **8**, 374–379.

Carvalho,P.C. *et al*. (2009) Charge prediction machine: tool for inferring precursor charge states of electron transfer dissociation tandem mass spectra. *Anal. Chem.*, **81**, 1996–2003.

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

Chen,L. and Yap,Y.L. (2008) Automated charge state determination of complex isotope-resolved mass spectra by peak-target Fourier transform. *J. Am. Soc. Mass Spectrom.*, **19**, 46–54.

Chi,A. *et al*. (2007) Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl Acad. Sci. USA*, **104**, 2193–2198.

Cociorva,D. *et al*. (2007) Validation of tandem mass spectrometry database search results using DTASelect. *Curr. Protoc. Bioinformatics*, Chapter 13, Unit 13.4.

Eng,J.K. *et al*. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.

Forbes,A.J. *et al*. (2001) Toward efficient analysis of >70 kDa proteins with 100% sequence coverage. *Proteomics*, **1**, 927–933.

Horn,D.M. *et al*. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.*, **11**, 320–332.

Makarov,A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.*, **72**, 1156–1162.

Mayampurath,A.M. *et al*. (2008) DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics*, **24**, 1021–1023.

McDonald,W.H. *et al*. (2004) MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.*, **18**, 2162–2168.

Perkins,D.N. *et al*. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Xu,T. *et al*. (2006) ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. *Mol. Cell. Proteomics*, **5**, S174.

YADA: A tool for taking the most out of high-resolution spectra

Paulo C Carvalho[1,2*], Tao Xu[2], Xuemei Han[2], Daniel Cociorva[2], Valmir C Barbosa[1], and John R Yates, III[2]

[1]Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Caixa Postal 68511, 21941-972 Rio de Janeiro - RJ, Brazil.

[2]Department of Chemical Physiology, The Scripps Research Institute, N Torrey Pines 10550, La Jolla, CA 92037, USA.

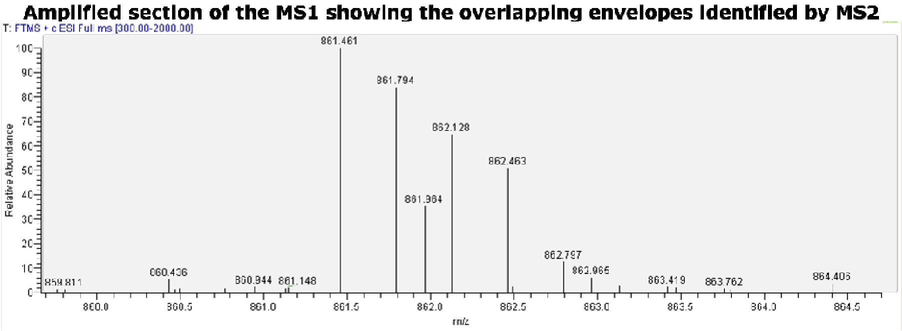
This text contains supplementary information to the manuscript YADA: A tools for taking the most out of high-resolution spectra.

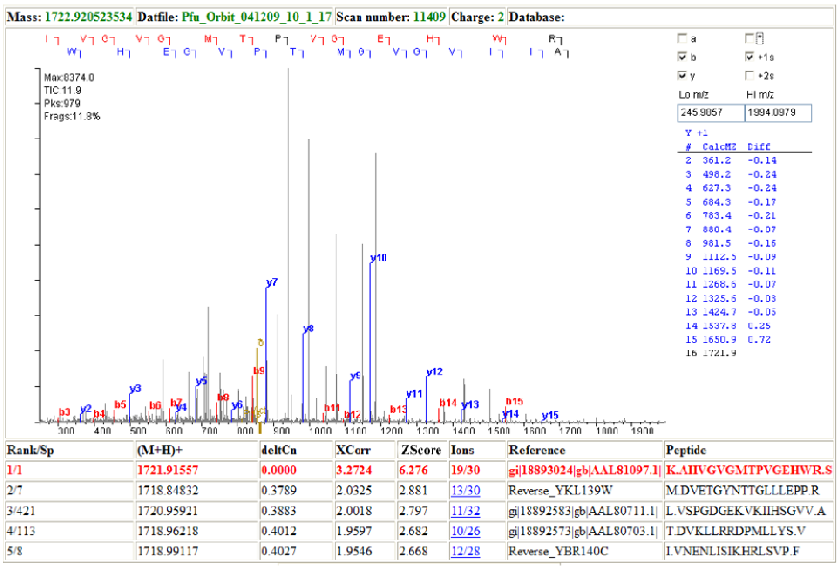
1 Description of the strategy used for false discovery rate estimation

Tandem mass spectra were extracted from the raw files; the spectra were searched against the yeast protein database plus common contaminant proteins; all sequences were downloaded as FASTA-formatted from the EBI-IPI protein database. To calculate confidence levels and false-positive rates, a decoy database that contained the reverse sequences of the original dataset appended to the target database was used (Peng et al., 2003), and the best matching sequences from the combined database were indicated by ProLuCID. The searches were done on a cluster of Intel Xeon 80 processors running the Linux operating system. The peptide mass search tolerance was set to 3 Da. No differential modifications were considered. For the aqueous digestion, the mass of the amino acid cysteine was statically modified by +57 Da due to the carboxyamidomethylation of the sample. No enzymatic cleavage conditions were imposed on the database search, so the search space included all candidate peptides whose theoretical mass fell within the 3 Da mass tolerance window, regardless of their tryptic status.

The validity of peptide / spectrum matches was assessed in DTASelect 2 (Tabb et al., 2002) according to the ProLuCID cross-correlation score (XCorr) and the ProLuCID normalized difference in cross-correlation score (DeltaCN). The search results were grouped by charge state (+1, +2, …, +21) and tryptic status (fully tryptic, half-tryptic, and non-tryptic), resulting in distinct subgroups. In each of the subgroups, the distribution of XCorr and DeltaCN values for the direct and decoy database hits was obtained, and the two subsets were separated by quadratic discriminant analysis. Outlier points in the two distributions (for example, matches with very low XCorr but very high DeltaCN) were discarded. Full separation of the direct and decoy subsets is not generally possible; therefore, the discriminant score was set such that a false-discovery rate of 1% was determined based on the number of accepted decoy database peptides. This procedure was independently performed on each data subset, resulting in a false-positive rate independent of tryptic status or charge state. In addition, a minimum sequence length of 7 amino-acid residues was required, and each protein on the list was supported by at least two peptide identifications. These additional requirements, especially the latter, resulted in the elimination of most decoy database and false-positive hits, as these tended to be overwhelmingly present as proteins identified by single peptide matches. After this last filtering step, the estimated false-discovery rate was reduced to below 1%.

Amplified section of the MS1 showing the overlapping envelopes identified by MS2





Mass: 1722.920523534 | **Datfile:** Pfu_Orbit_041209_10_1_17 | **Scan number:** 11409 | **Charge:** 2 | **Database:**

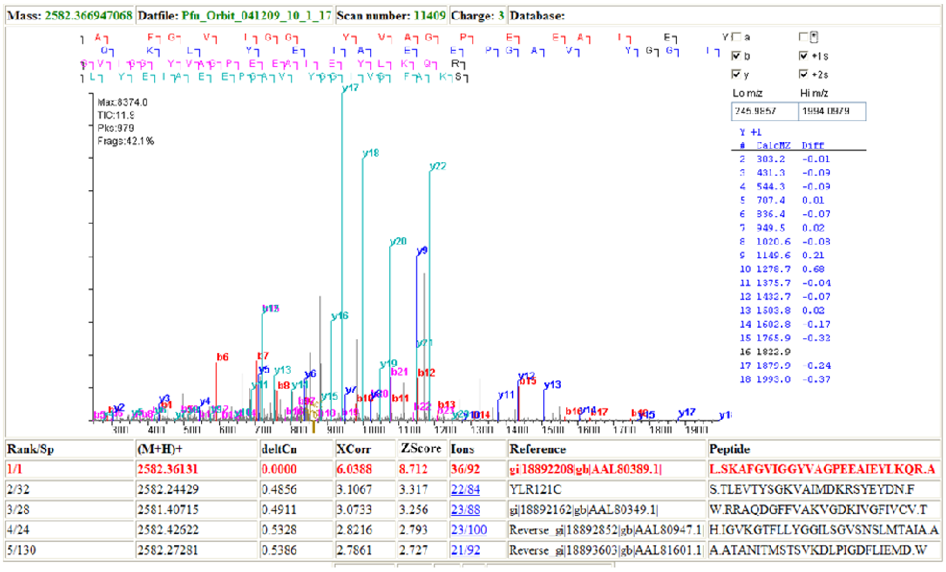
□ a □
☑ b ☑ +1s
☑ y □ +2s

Lo m/z 245.9057 Hi m/z 1994.0979

```
Y +1
#  CalcMZ  Diff
2  361.2   -0.14
3  498.2   -0.24
4  627.3   -0.24
5  684.3   -0.17
6  783.4   -0.21
7  880.4   -0.07
8  961.5   -0.16
9  1112.5  -0.09
10 1169.5  -0.11
11 1268.6  -0.07
12 1325.6  -0.03
13 1424.7  -0.05
14 1537.8  0.25
15 1650.9  0.72
16 1721.9
```

Max:8374.0
TIC:11.9
Pks:979
Frags:11.8%

Rank/Sp	(M+H)+	deltCn	XCorr	ZScore	Ions	Reference	Peptide
1/1	1721.91557	0.0000	3.2724	6.276	19/30	gi|18893024|gb|AAL81097.1|	K.AIIVGVGMTPVGEHWR.S
2/7	1718.84832	0.3789	2.0325	2.881	13/30	Reverse_YKL139W	M.DVETGYNTTGLLLEPP.R
3/421	1720.95921	0.3883	2.0018	2.797	11/32	gi|18892583|gb|AAL80711.1|	L.VSPGDGEKVKIIHSGVV.A
4/113	1718.96218	0.4012	1.9597	2.682	10/26	gi|18892573|gb|AAL80703.1|	T.DVKLLRRDPMLLYS.V
5/8	1718.99117	0.4027	1.9546	2.668	12/28	Reverse_YBR140C	I.VNENLISIKHRLSVP.F



Mass: 2582.366947068 | **Datfile:** Pfu_Orbit_041209_10_1_17 | **Scan number:** 11409 | **Charge:** 3 | **Database:**

□ a □
☑ b ☑ +1s
☑ y ☑ +2s

Lo m/z 245.9857 Hi m/z 1994.0979

```
Y +1
#  CalcMZ  Diff
2  303.2   -0.01
3  431.3   -0.09
4  544.3   -0.09
5  707.4   0.01
6  836.4   -0.07
7  949.5   0.02
8  1020.6  -0.03
9  1149.6  0.21
10 1278.7  0.68
11 1375.7  -0.04
12 1432.7  -0.07
13 1503.8  0.02
14 1602.8  -0.17
15 1765.9  -0.32
16 1822.9
17 1879.9  -0.24
18 1993.0  -0.37
```

Max:8374.0
TIC:11.9
Pks:979
Frags:42.1%

Rank/Sp	(M+H)+	deltCn	XCorr	ZScore	Ions	Reference	Peptide
1/1	2582.36131	0.0000	6.0388	8.712	36/92	gi|18892208|gb|AAL80389.1|	L.SKAFGVIGGYVAGPEEAIEYLKQR.A
2/32	2582.24429	0.4856	3.1067	3.317	22/84	YLR121C	S.TLEVTYSGKVAIMDKRSYEYDN.F
3/28	2581.40715	0.4911	3.0733	3.256	23/88	gi|18892162|gb|AAL80349.1|	W.RRAQDGFFVAKVGDKIVGFIVCV.T
4/24	2582.42622	0.5328	2.8216	2.793	23/100	Reverse_gi|18892852|gb|AAL80947.1|	H.IGVKGTFLLYGGILSGVSNSLMTAIA.A
5/130	2582.27281	0.5386	2.7861	2.727	21/92	Reverse_gi|18893603|gb|AAL81601.1|	A.ATANITMSTSVKDLPIGDFLIEMD.W

Figure S1: The top panel shows a zoomed-in view of a region of an MS1 spectrum containing overlapping isotopic envelopes. The two lower panels show the corresponding tandem mass spectra. As can be noted, two distinct peptides were confidently identified with high ZScore's, as shown in red.

ANEXO X

PubMed

U.S. National Library of Medicine
National Institutes of Health

Display Settings: Abstract

Bioinformatics. 2010 Mar 15;26(6):847-8. Epub 2010 Jan 26.

XDIA: improving on the label-free data-independent analysis.

Carvalho PC, Han X, Xu T, Cociorva D, Carvalho Mda G, Barbosa VC, Yates JR 3rd.

Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Caixa Postal 68511, 21941-972 Rio de Janeiro, Brazil. paulo@pcarvalho.com

SUMMARY: XDIA is a computational strategy for analyzing multiplexed spectra acquired using electron transfer dissociation and collision-activated dissociation; it significantly increases identified spectra (approximately 250%) and unique peptides (approximately 30%) when compared with the data-dependent ETCaD analysis on middle-down, single-phase shotgun proteomic analysis. Increasing identified spectra and peptides improves quantitation statistics confidence and protein coverage, respectively. AVAILABILITY: The software and data produced in this work are freely available for academic use at http://fields.scripps.edu/XDIA CONTACT: paulo@pcarvalho.com SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics online.

PMID: 20106817 [PubMed - in process]

PMCID: PMC2832823 [Available on 2011/3/15]

Publication Types, Grant Support

LinkOut - more resources

You are here: NCBI > Literature > PubMed

Write to the Help Desk