SEQUENCIAMENTO *DE NOVO* DE PEPTÍDEOS UTILIZANDO GRAFOS PARA ESPECTROS DE MASSA MULTIPLEX

José Cláudio Garcia Damaso

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

Orientadora: Beatriz de Souza Leite Pires de Lima

Rio de Janeiro

Abril de 2017

SEQUENCIAMENTO *DE NOVO* DE PEPTÍDEOS UTILIZANDO GRAFOS PARA ESPECTROS DE MASSA MULTIPLEX

José Cláudio Garcia Damaso

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 CIVIL.

Examinada por:

_____

Prof[a]. Beatriz de Souza Leite Pires de Lima, D.Sc.

_____

Prof[a]. Solange Guimarães, D.Sc.

_____

Prof. Fábio Cesar Souza Nogueira, D.Sc.

_____

Prof. Floriano Paes Silva Júnior, D.Sc.

_____

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

_____

Prof. Rafael Donádelli Melani, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

ABRIL DE 2017

Damaso, José Cláudio Garcia

Sequenciamento *De Novo* de Peptídeos Utilizando Grafos para Espectros de Massa Multiplex/ José Cláudio Garcia Damaso. – Rio de Janeiro: UFRJ/COPPE, 2017

XVII, 119 p.: il.; 29,7 cm.

Orientadora: Beatriz de Souza Leite Pires de Lima

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Civil, 2017.

Referências Bibliográficas: p. 115-119.

1. Proteômica. 2. Espectro de massa. 3. Espectro multiplex. 4. Sequenciamento *de novo* de peptídeo. I. Lima, Beatriz de Souza Leite Pires de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título

"Eu tenho cada vez menos tempo, embora tenha cada vez mais coisas para dizer. E o que eu tenho para dizer é, mais e mais, algo que se move para frente, junto ao movimento de meu pensar."

Pablo Picasso

À minha avó materna, Maria (*in memorian*)

AGRADECIMENTOS

Primeiramente agradeço a Deus, a Luz, a Força Superior, que oportunizou minha vida e permitiu todas as experiências que tenho vivido.

Às professoras Beatriz e Solange pela orientação e incentivo para a realização deste trabalho, principalmente Solange, que esteve sempre presente e ao meu lado, não me permitindo errar e nem desviar do caminho, assistiu minhas angústias bem de perto, foi sempre extremamente caridosa em minhas fraquezas, com quem eu dividi muitas alegrias e tristezas, que tem minha eterna gratidão.

Aos meus pais, Arthur e Nolmira, por tudo que fazem, por todo o carinho, amor e paciência, pelo exemplo.

À Hanriette e Sofia, que acompanharam de perto minha luta, por permitirem minha ausência devido às horas dedicadas à construção desse trabalho.

À minha avó materna, Maria (*in memorian*), por ter acreditado e incentivado este título ainda em meus tempos de criança, mesmo sem saber.

Aos meus irmãos, Fábio, Kátia e Lúcia, por participarem desde sempre de minha formação e aceitarem meu distanciamento temporário para realização dessa jornada.

À Lúcio, humano, amigo e principal incentivador desta etapa, pelas horas de conversas, preocupação e ajuda direta no desenvolvimento desta tese.

Aos professores Aurélio, Elaine, Pauli e Selma, pelo interesse e construção das ideias que somaram no desenvolvimento desse trabalho.

A todos os professores de meu departamento (UFF-VCO), que me ajudaram, direta ou indiretamente, principalmente o prof. Arlindo, sereno e justo, agradeço por acreditarem e contribuírem para a realização desse objetivo.

À equipe do laboratório de computação da Engenharia Civil, Orlando, Célio e Amanda, pelo apoio operacional em diversas oportunidades.

Ao professor Gilberto Domont, Magno Junqueira, Gabriel, Fábio e Rafael, todos do Laboratório de Espectrometria de Massa do Instituto de Química da UFRJ, por terem sido tão prestativos e amigos ao longo de todo meu doutoramento.

A todos aqueles que contribuíram na elaboração deste trabalho.

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

SEQUENCIAMENTO *DE NOVO* DE PEPTÍDEOS UTILIZANDO GRAFOS PARA ESPECTROS DE MASSA MULTIPLEX

José Cláudio Garcia Damaso

Abril/2017

Orientadora: Beatriz de Souza Leite Pires de Lima

Programa: Engenharia Civil

Esta tese apresenta uma abordagem para sequenciamento *de novo* (*in silico*) de peptídeos em espectros MS^2 multiplex, adquiridos em espectrômetros de massa, que contêm fragmentos de mais de um peptídeo na mesma janela de fragmentação. Foi desenvolvido um método simples e expedito, denominado DNbuilder, para o sequenciamento *de novo* de peptídeos com uma pontuação que considera as intensidades dos fragmentos do tipo **y**. O problema foi modelado através de grafos e adotou-se o algoritmo de busca DFS (*Depth-first search*) para se obter as sequências candidatas dos peptídeos. Identificadas as massas sobre cargas (*m/z*) dos íons peptídeos monoisotópicos de carga +2 presentes na janela selecionada para fragmentação do primeiro espectro de massa, MS^1. A metodologia multiplex fundamenta-se na alteração das intensidades dos picos fragmentos do segundo espectro, MS^2, para cada novo peptídeo a ser sequenciado da janela, podendo incluir a retirada, ou atenuação, de picos do espectro correspondentes a fragmentos do tipo y ou b dos peptídeos já identificados. Os programas de sequenciamento *de novo* usados para validar programa DNbuilder e a metodologia multiplex foram o Peaks 8, pNovoPlus e Novor 1.3.489. Espectros de uma amostra de tireoide adquiridos em janelas de 20 *m/z* foram usados nos testes de avaliação da metodologia multiplex. Os resultados mostram que a metodologia, mesmo que simples, melhora o sequenciamento *de novo* dos peptídeos presentes nos espectros multiplex MS^2, aumentando o número de resíduos de aminoácidos corretamente posicionados nas sequências encontradas, mostrando que há um caminho possível para o sequenciamento *de novo* de peptídeos em espectros multiplex para janelas amplas.

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

DE NOVO PEPTIDE SEQUENCING USING GRAPHS FOR MULTIPLEX MASS SPECTRUM

José Cláudio Garcia Damaso

April/2017

Advisor: Beatriz de Souza Leite Pires de Lima

Department: Civil Engineering

This thesis presents an approach for *de novo* peptide sequencing (*in silico*) using multiplex MS^2 mass spectra, acquired by a mass spectrometer, containing fragments of more than one peptide in the same MS^2 spectrum. A program, named DNbuilder, was developed for *de novo* peptide sequencing, in which the cost function considers y-fragments intensity. The *de novo* peptide sequencing problem was modeled in graphs using Depth-first search, DFS, as the search algorithm for peptide sequencing over the graph. Within the predefined fragmentation window, monoisotopic charged +2 peptide ions were identified and selected as the MS^2 targets. The multiplex methodology consists of the intensity changes of selected peaks from the MS^2 multiplex spectrum, including the elimination or reduction of the y or b fragments of the previously identified peptides present in the MS^2 multiplex window. The state-of-the-art programs used in the multiplex *de novo* peptide sequencing tests were Peaks 8, pNovoPlus e Novor 1.3.489, as well as, DNbuilder. A set of spectra from a thyroid sample acquired in 20 *m/z* window size was used to evaluate the multiplex methodology. The results show the methodology, even if simple, increased the number of correct identified peptide amino acids in multiplex spectra, an evidence that there are ways of *de novo* sequencing multiplex spectra.

Índice

Lista de Figuras

Lista de Tabelas

Relação de Símbolos e Abreviaturas

AC – Corrente alternada; do inglês, *alternating corrent*

ACN – Acetonitrila

C – Símbolo do elemento químico carbono

Cα – Carbono alfa

CID – Dissociação Induzida por Colisão; do inglês, *Colision-Induced Dissociation*

COOH – Grupo carboxil

Da – Dalton, unidade de massa que é numericamente igual à massa atômica unificada

Da/z – Unidade de medida de massa sobre carga

DDA – Aquisição dependente de dados; do inglês, *Data Dependent Aquisition*

DFS – Busca em profundidade; do inglês, *Depth-First Search*

DIA – Aquisição Independente de Dados; do inglês, *Data Independent Analysis*

DTT – Ditiotreitol

ECD – Dissociação por Captura de Elétrons; do inglês, *Electron Capture Dissociation*

EDD – Dissociação por Separação de Elétrons; do inglês, *Electron Detachment Dissociation*

ESI – Ionização por Eletrospray; do inglês, *Electrospray Ionization*

ETD – Dissociação por Transferência de Elétrons; do inglês, *Electron Transfer Dissociation*

HCD – Dissociação por Maior Energia de Colisão; di inglês, *Higher-energy Collisional Dissociation*

HPLC – Cromatografia Líquida de Alta Eficiência; do inglês, *High-Performance Liquid Chromatography*

I – Intensidade de um pico do espectro

IAA – Iodocetamida

MALDI – Ionização por Dessorção a *Laser* assistida por Matriz; do inglês, *Matrix-Assisted Laser Desorption/Ionization*

MGF – extensão de formato de arquivo usado em espectrometria de massa pelo programa para identificação Mascot; do inglês, *Mascot Generic Format*

MS2 – extensão de formato de arquivo usado em espectrometria de massa

MS^1 ou MS – Espectro de massa dos íons totais; do inglês, *Mass spectrum*

MS^2 ou MS/MS – Espectro de massa dos íons-fragmentos

N – Símbolo do elemento químico nitrogênio

NAD – Número de Aresta Dupla

NH_3^+ – Grupo amino
Nr – Número total de resíduos do peptídeo

PC – Pontuação Final do Caminho

PPM – Parte por Milhão

PSM – Espectro-peptídeo correspondente; do inglês, *Peptide Spectrum Match*

R – Cadeia lateral de um aminoácido

RAW – extensão de formato de arquivo binário usado em espectrometria de massa por fabricantes de espectrômetros de massa.

RF – Radiofrequência

TFA – Trifluoroacético

Th – Thonson, unidade de medida de massa

ToF – Tempo de Vôo; do ingles, *Time of Flight*

z – Carga de um íon

δ_{max} – Tolerância máxima entre a massa calculada e a massa teórica de um aminoácido

1 INTRODUÇÃO

A proteômica é um conjunto de técnicas de análise para o estudo de proteínas em larga escala. Proteômica é uma disciplina da era pós-genômica que abrange a identificação e a quantificação das proteínas de um proteoma, ou seja, um conjunto de proteínas expressas por um genoma. A proteômica engloba os processos tanto de identificação quanto de quantificação dos componentes proteicos de um proteoma, tais como estruturas primárias e covalentes, conformação, modificações pós-traducionais, diferenciações de isoformas, polimorfismo, *splicing*, localização, etc.

A análise de um proteoma abrange uma série de técnicas físico-químicas e computacionais que são empregadas para qualificar e quantificar um conjunto de proteínas expressas em determinadas condições. A identificação das proteínas presentes em uma amostra biológica é feita por espectrometria de massa [1].

O espectrômetro de massa é um equipamento analisa a relação massa/carga (*m/z*) de um íon. Essa relação *m/z*, bem como sua intensidade, é registrada em espectros de massa. A partir dos espectros de massa registrados pelo espectrômetro é possível identificar as proteínas. Esta identificação faz uso de técnicas computacionais que incluem buscas em bancos de dados de genomas e transcriptomas que contenham sequências de proteínas conhecidas. Os espectros de massa usados nas identificações em bancos de dados também possibilitam simples reconstrução de sequências desconhecidas para ajudar na identificação de proteínas por alinhamento de sequência e similaridade. Todo o processo de interpretação de resultados está associado direta ou indiretamente à ciência da computação, em função da imensa quantidade de dados (espectros) gerados em uma única corrida do espectrômetro MS^2.

Neste trabalho abordamos as técnicas usadas na reconstrução da sequência de ácidos aminados de peptídeos usando somente o espectro de massa dos seus íons fragmentos, processo conhecido como sequenciamento *de novo*. O objetivo é o desenvolvimento de uma metodologia que permita identificar proteínas independentemente da existência de genomas e transcriptomas conhecidos ou diretamente da sequência de proteínas depositadas em banco de dados. O foco passa a ser a reconstrução da sequência peptídica para identificar proteínas por similaridade de sequência.

1.1 Motivação

O sequenciamento *de novo* de peptídeos, apesar de parecer um problema de solução simples, ainda não tem resultados plenamente satisfatórios, já que é muito dependente da boa qualidade dos espectros de massa adquiridos pelos espectrômetros. A importância do sequenciamento *de novo*, no entanto, fica clara quando a proteína procurada não está presente nos bancos de dados, impossibilitando a identificação da proteína por comparação PSM, *Peptide-Spectrum Match*. Com o desenvolvimento de novos equipamentos, e consequentemente de novas técnicas de análise mais ousadas na espectrometria de massas, gerando espectros cada vez mais complexos, cria-se a necessidade de acompanhar este desenvolvimento com técnicas numéricas capazes de analisar e interpretar estes dados. A curiosidade sobre a possibilidade de produzir sequenciamento *de novo* nestes espectros mais complexos, assim como a vontade de melhorar os resultados para o problema de sequenciamento *de novo* usando espectros MS^2, foram as fontes de motivação para este estudo.

1.2 Objetivos

A existência de espectros de massa MS^2 que registram íons-fragmentos de mais de um peptídeo não é incomum e constitui um dos desafios para o sequenciamento *de novo* de peptídeos. Assim, o objetivo deste trabalho é propor uma solução para sequenciar a maior quantidade possível de peptídeos[1] nos espectros MS^2 ditos multiplex. A solução perfeita é aquela que consegue sequenciar todos os peptídeos presentes nos espectros MS^2 adquiridos em espectrômetros de massa a partir das amostras de peptídeos ali inseridas.

A solução ideal para o problema de sequenciamento *de novo* é aquela que consegue sequenciar a maior quantidade possível de peptídeos que foram fragmentados e registrados num mesmo espectro de massa, ou mesmo, diante das dificuldades existentes e descritas neste trabalho, aumentar a quantidade de resíduos de aminoácidos corretamente sequenciados destes peptídeos presentes.

Para se alcançar o objetivo desse trabalho de aumentar a quantidade dos sequenciamentos *de novo* realizados, há que se atingirem as seguintes etapas:

[1] Foram considerados somente os peptídeos de carga +2 presentes no MS^1, que geram fragmentos de carga +1 nos espectros MS^2, devido à dificuldade em deconvoluir os fragmentos de carga mais alta nestes espectros adquiridos em Orbitraps.

- ✓ Desenvolver uma metodologia própria para o sequenciamento *de novo* de peptídeos possibilitando o uso de diferentes estratégias, tanto na reconstrução das sequências a partir dos picos dos espectros, quanto na heurística de pontuação usada para se estabelecer as melhores sequências dentro de uma lista de possíveis candidatas. A programação decorrente foi designada DNBuilder;

- ✓ Implementação da visualização dos peptídeos presentes nos espectros MS^1 de janela ampla dentro da metodologia computacional para o sequenciamento *de novo*;

- ✓ Identificar os programas mais usados para o sequenciamento *de novo* de peptídeos dentro da literatura, para testes comparativos;

- ✓ Criar uma estratégia de estudo e análise dos espectros multiplex;

- ✓ Por causa das diferentes intensidades dos peptídeos ionizados registradas nos espectros MS^1, há a necessidade de determinar um mecanismo de descomplexação do espectro MS^2 através da retirada ou alteração das intensidades dos picos do peptídeo mais intenso do espectro MS^2 para que os picos dos ions-fragmentos dos peptídeos menos intensos tornem-se visíveis para as heurísticas de sequenciamento *de novo*,

- ✓ Comparar os resultados obtidos pelos diferentes programas usados nesta tese, incluindo o DNbuilder.

1.3 Contribuições

Este trabalho faz uso de técnicas já conhecidas nas soluções para o problema de sequenciamento *de novo*, como grafo. Também faz uso da junção de técnicas, como a busca DFS implementada usando a Programação Dinâmica, que nada mais é do que uma programação recursiva, que retorna uma lista de soluções viáveis para a sequência do peptídeo. No entanto, o estudo do sequenciamento *de novo* em espectros multiplex usando janelas de fragmentação amplas, que levam a um registro complexo de íons-

fragmentos nos espectros de massa, não foi ainda estudado. Um trabalho recente [2] aborda uma descomplexação dos espectros de peptídeos isolados em janelas usuais de 2 a 3 *m/z*, que inadvertidamente contenham outros peptídeos ali co-fragmentados e registrados. Estes espectros são de menor complexidade que os de janela ampla aqui abordados. Assim, o conjunto do trabalho traz novos estudos e abordagens para o problema de sequenciamento *de novo* indicando um caminho computacional para a descomplexação dos espectros de massa em espectros específicos para cada peptídeo a ser sequenciado.

1.4 Organização dos capítulos

O presente texto está dividido em 7 capítulos. O Capítulo 1 faz uma breve introdução desta tese e descreve as motivações, objetivos e contribuições deste trabalho. No Capítulo 2 é feita uma breve definição de proteína, para que o leitor que não tenha intimidade com a área biológica possa compreender o problema. O Capítulo 3 trata brevemente da espectrometria de massa, com o intuito de introduzir o leitor na compreensão do problema numérico do sequenciamento *de novo* de peptídeos. O Capítulo 4 aborda o sequenciamento *de novo* e as características do processo. Neste capítulo são abordadas as várias heurísticas até agora usadas por diversas ferramentas para resolver o problema de sequenciamento *de novo*, desde a seleção dos sinais ou picos relevantes dos espectros, as opções de modelagem através de grafos ou combinação de sequências, os critérios de busca das sequências usando custos em nós e/ou arestas, até a escolha dos algoritmos de busca das melhores sequências. No Capítulo 5 é apresentada a heurística do sequenciamento *de novo* desenvolvida neste trabalho, nomeada DNbuilder. O Capítulo 6 mostra os resultados das metodologias e programas usados na tese. Por fim, o Capítulo 7 apresenta as conclusões do trabalho e propostas futuras.

2 PROTEÍNAS

As proteínas são macromoléculas que desempenham funções específicas nos seres vivos. São compostas por uma sequência de aminoácidos ligados covalentemente entre si. Na natureza há uma quantidade limitada de aminoácidos diferentes que podem fazer parte da composição de uma proteína.

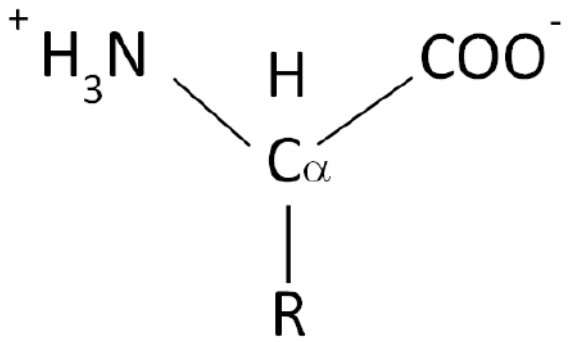


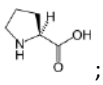
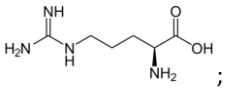
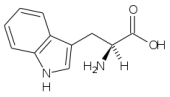
Figura 1: Estrutura de um aminoácido genérico em meio fisiológico

Um aminoácido é composto de um carbono alfa (Cα), ligado a um grupo amino (NH_2), a um grupo carboxila (COOH), um hidrogênio e a uma cadeia lateral que define o aminoácido. Em meio fisiológico o grupo carboxil desprotona e o amino é protonado, como apresentado na Figura 1 que mostra esquematicamente a estrutura de um aminoácido em meio fisiológico, onde R representa a cadeia lateral. A Tabela 1 apresenta a lista dos 20 aminoácidos mais comuns, a nomenclatura de uma letra internacionalmente adotada e a estrutura da cadeia lateral.

Tabela 1: Nome do aminoácido, simbologia com 3 e 1 letras e composição da cadeia lateral dos aminoácidos [60]

Aminoácido	Símbolo		Estruturas(*)
Glicina	Gly	G	**H**- CH (NH_2) – COOH
Alanina	Ala	A	**CH_3**- CH (NH_2) – COOH
Serina	Ser	S	**OH-CH_2**- CH (NH_2)- COOH
Prolina	Pro	P	**CH_2-CH_2-CH_2**-CH (NH_2)- COOH(1)
Valina	Val	V	**CH_3-CH(CH_3)**-CH (NH_2)- COOH
Treonina	Thr	T	**OH-CH (CH_3)**- CH (NH_2)- COOH
Cisteína	Cys	C	**SH-CH_2**- CH (NH_2)- COOH
Leucina	Leu	L	**CH_3(CH_2)$_3$-CH2**-CH (NH_2)- COOH
Isoleucina	Ile	I	**CH_3-CH_2-CH (CH_3)**-CH (NH_2)- COOH
Aspargina	Asn	N	**NH_2-CO-CH_2**- CH (NH_2)- COOH
Ácido aspártico	Asp	D	**HCOO-CH_2**- CH (NH_2)- COOH
Glutamina	Gln	Q	**NH_2-CO-CH_2-CH_2**- CH (NH_2)- COOH
Lisina	Lys	K	**NH_3-CH_2-CH_2-CH_2-CH_2**- CH (NH_2)- COOH
Ácido glutâmico	Glu	E	**HCOO-CH_2-CH_2**- CH (NH_2)- COOH
Metionina	Met	M	**CH_3-S-CH_2-CH_2**- CH (NH_2)- COOH
Histidina	His	H	**H-($C_3H_2N_2$)-CH_2**- CH (NH_2)- COOH
Fenialanina	Phe	F	**C_6H_5-CH_2**-CH (NH_2)- COOH
Arginina	Arg	R	**C(NH_2)-NH-CH_2-CH_2-CH_2**- CH (NH_2)- COOH(2)
Tirosina	Tyr	Y	**OH-C_6H_4-CH_2**- CH (NH_2)- COOH
Triptofano	Trp	W	**R aromático**- CH (NH_2)- COOH (3)

(*) Cadeia lateral em negrito;

(1)prolina  ; (2) arginina  ; (3) triptofano 

As proteínas são formadas por uma sequência de aminoácidos unidos por ligações peptídicas que se formam quando a carboxila de um aminoácido se condensa ao grupo amino de outro, como esquematizado na Figura 2. É chamada de estrutura primária a sequência de aminoácidos listadas a partir do grupo amino, N-terminal, até o grupo carboxílico, C-terminal, como indicado na Figura 3. Um peptídeo é uma sequência de alguns aminoácidos até um limite informalmente aceito da massa molecular da insulina, em torno de 50 resíduos, ou seja, um pedaço menor da cadeia polipeptídica de uma proteína, como ilustrado na Figura 2.

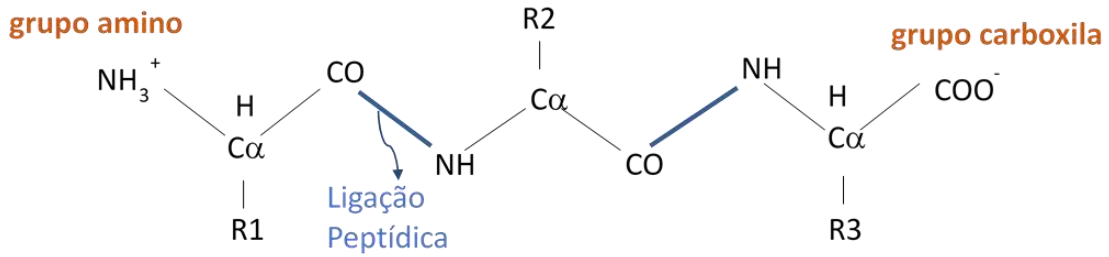


Figura 2: Peptídeo de três resíduos genéricos de aminoácidos

Os aminoácidos, do momento em que formam a cadeia polipeptídica, são denominados resíduos de aminoácidos, ou simplesmente resíduos, como esquematizado na Figura 3. Seu conceito pode ser definido como a massa de um aminoácido pertencente a uma proteína menos a perda decorrente da geração do peptídeo [40]. A cadeia polipeptídica feita pelas ligações dos átomos ...N-Cα-C-N-Cα-C... é chamada de cadeia principal ou esqueleto da proteína, em contraponto às cadeias laterais.

As massas teóricas dos resíduos serão dadas mais adiante, quando forem abordados os tipos de fragmentações geradas nos espectrômetros.

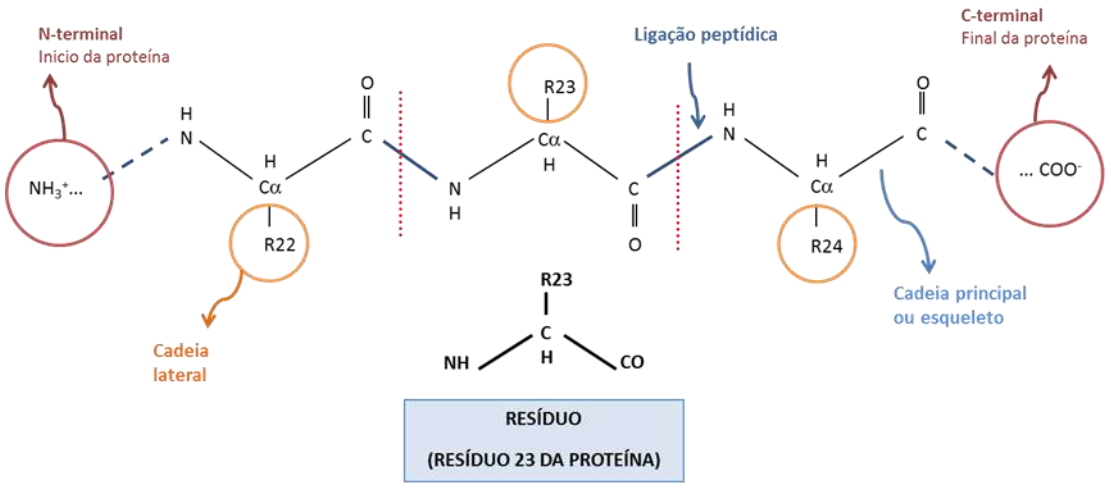


Figura 3: Esquema de uma proteína genérica

3 ESPECTROMETRIA DE MASSA

O espectrômetro de massa data do início do século XIX, mas o desenvolvimento da técnica de ionização por *electrospray* apresentada por FENN *et* al [3] impulsionou a utilização dos espectrômetros de massa nos estudos de moléculas biológicas de proteínas. Isso fez com que o sequenciamento *de novo* químico de peptídeos por degradação de Edman, fosse substituído pela espectrometria de massa por ser mais sensível e rápida. Além disso, a técnica não requer proteínas ou peptídeos purificados e pode trabalhar com proteínas modificadas [4].

Os espectrômetros de massa são construídos a partir de uma combinação de componentes para ionização, análise e detecção de massas. As fontes de ionizações utilizadas na proteômica são a ESI-*Electrospray Ionization* e MALDI-*Matrix-Assisted Laser Desorption/Ionization*, que promovem uma ionização suave sem fragmentar o peptídeo. Apenas os peptídeos ionizados são analisados pelo espectrômetro. As análises das massas podem ser feitas por *Quadupole*, ToF-*Time-of-Flight*, *Orbitrap*, *IonTrap* [5] e outros, e têm como função separar/isolar os peptídeos ionizados em função da relação *m/z*. Uma vez separados os íons são detectados e suas intensidades registradas. Usando programas operacionais, o espectrômetro de massa isola e fragmenta um íon, analisando os fragmentos detectados, num processo sequencial. A fragmentação pode ser feita em meio gasoso por dissociação de baixa energia, conhecida como *collision-induced dissociation* (CID), ou por dissociação em alta energia, *higher-energy collisional dissociation* (HCD), ou por dissociação por transferência de elétrons, *electron-transfer dissociation* (ETD), cuja energia de dissociação é escolhida pelo operador, e ainda por *electron capture dissociation* (ECD). Cada técnica de fragmentação produz fragmentos com características distintas. Os tipos de fragmentações CID e HCD são apropriados para análises *bottom-up*[2], pois não conseguem quebrar moléculas muito grandes, sendo que, dependendo do analisador, a fragmentação por HCD registra mais íons nas regiões de massas menores [6], o que o CID não faz. A abordagem *top-down*[3] é mais indicada em análise de proteína inteira, ou moléculas maiores.

Vale acrescentar aqui a importância da resolução e acurácia para o espectrômetro de massa. A utilização de analisadores de alta resolução permite obter

[2] Técnica proteômica que utilize uma digestão enzimática da proteína antes da análise por espectrometria de massa.
[3] Técnica proteômica que analisa a proteína intacta por espectrometria de massa.

massa mais precisa da amostra, resultado da combinação da focalização por energia e por momento [7], acoplado a componentes eletrostáticos e magnéticos (focalização dupla) [8]. Quanto maior a resolução (*high-resolution*) de um analisador, mais precisa será a massa da amostra analisada, chegando a mais de quatro casas decimais. A resolução alta permite separar íons com *m/z* próximas como por exemplo íons isótopos muito carregados. A acurácia de massas é um parâmetro importante de um espectrômetro de massas e determina a especificidade da medida de massa [9]. A definição mais comumente empregada atualmente para acurácia de massas leva em consideração a diferença entre a massa exata (teórica) e a massa medida (calculada) em relação à massa teórica, expressa em ppm-partes por milhão. A acurácia é aferida através de amostras padrão onde os picos de *m/z* são conhecidos e podem ser alinhados com a aquisição do espectrômetro. Entretanto, como essa é uma unidade relativa, pode variar significativamente com a massa do íon em análise [10]. Espectros provenientes de equipamentos com maior acurácia e resolução oferecem dados de melhor qualidade para o sequenciamento *de novo* [11].

O espectrômetro de massa pode estar conectado em uma cromatografia líquida, como por exemplo, uma nano HPLC (*high-performance liquid chromatography*), que separa os peptídeos dependendo do tipo de matriz usado na coluna. A Figura 4, foto retirada pelo autor desta tese, mostra uma nano HPLC acoplada a um espectrômetro híbrido da Thermo[4], utilizado para efetuar experimentos em proteômica, dotado de sistema um *electrospray* que ioniza os peptídeos, contém dois tipos de analisadores, quais sejam, Orbitrap e Linear Trap, e também contém métodos de fragmentação CID, ETD e HCD. Este equipamento encontra-se na Unidade de Proteômica do Instituto de Química da Universidade Federal do Rio de Janeiro.

A técnica de análise feita por espectrometria de massa pode ser do tipo *top-down, middle-down* ou *bottom-up.* A análise *top-down* consiste na injeção da amostra de proteínas intactas no espectrômetro. A análise *middle-down*, injeção de peptídeos muito grandes. A análise mais comum é do tipo *bottom-up*, abordada neste trabalho, que vem a ser a inferência de proteínas a partir de seus peptídeos obtidos por hidrólise química ou enzimática. Se a mistura proteica não for complexa, utiliza-se a hidrólise total; se complexa, a mistura é pré-fracionada e cada fração hidrolisada separadamente. Em

[4] Thermo Fisher Scientific, Inc., USA. Empresa multinacional fornecedora de equipamentos científicos

ambos os casos os peptídeos presentes nos hidrolizados são levados à cromatografia líquida para fracionamento e injeção no espectrômetro de massa.

A enzima comumente usada para digestão é a tripsina. A hidrólise tríptica corta ligações peptídicas após a lisina(K) e arginina(R), desde que não sejam seguidas pela prolina (P), i.e., ligações K-P ou R-P. Este procedimento gera conjuntos de peptídeos de uma ou mais proteínas da amostra. Os peptídeos obtidos contêm, na sua maioria, uma lisina ou uma arginina como resíduo terminal, além de peptídeos N- e C-terminais e os provenientes de clivagens inespecíficas ou de reações químicas espúrias.

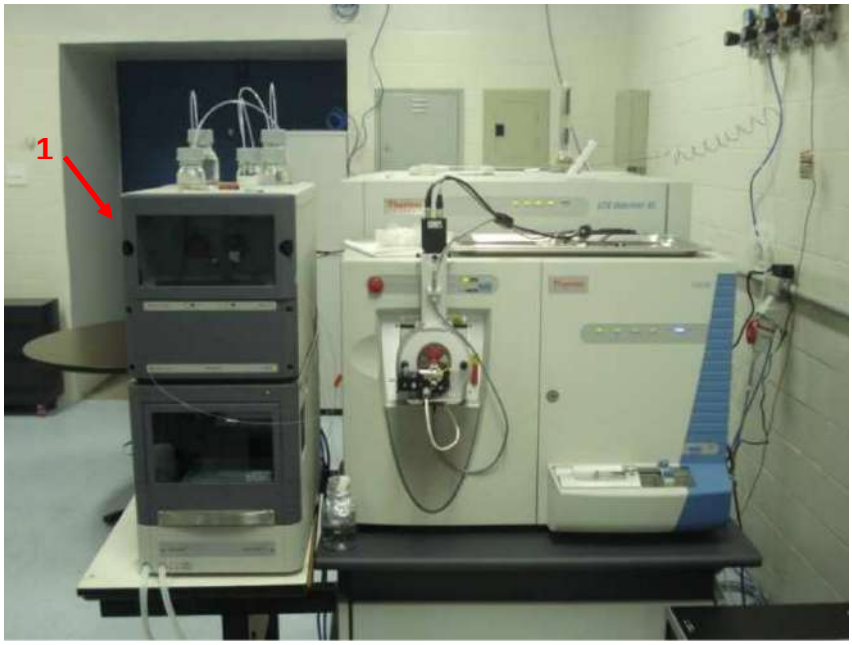


Figura 4: Espectrômetro de massa híbrido utilizado para efetuar experimentos em proteômica. 1) Nano HPLC; Na parte posterior e anterior encontram-se os analisadores Orbitrap XL e Linear Trap LTQ. As possibilidades de fragmentação ligadas aos analisadores são CID, ETD e HCD

O processo usual para adquirir os espectros de massa consiste em injetar a amostra na coluna da fase reversa da cromatografia líquida, que separa os peptídeos de acordo com sua hidrofobicidade e está associada em série ao espectrômetro. Saindo da coluna a amostra segue por um sistema de ionização ao ser injetado no espectrômetro, o *electrospray*.

Existe uma diferença de potencial aplicada na entrada do espectrômetro, onde se situa o *electrospray*. A fonte ESI ioniza e auxilia na vaporização do solvente em que está o peptídeo para que possa ser detectado e analisado. Peptídeos não ionizados não são detectados no espectrômetro de massa.

Os espectrômetros registram as massas/cargas de cada peptídeo da amostra com uma intensidade que é proporcional à quantidade do peptídeo que foi ionizado. Se o peptídeo está presente na amostra, mas não ioniza, ele não vai ser detectado.

3.1 Espectro de massa

Um espectro de massa é o registro da intensidade, abundância e razão da massa sobre a carga dos peptídeos ou fragmentos. O equipamento detecta a intensidade (*I*) dos peptídeos ou dos fragmentos de peptídeos. Por simplificação *m/z* será chamada de relação massa-carga. A intensidade está associada à quantidade de íons de mesma massa-carga que foram ionizados e foram detectados e não à quantidade deles na amostra. A Figura 5 mostra um exemplo de espectro de massa dos fragmentos de um peptídeo, ou MS^2, plotado em gráfico, com o registro das intensidades no eixo das ordenadas e a *m/z* no eixo das abscissas.

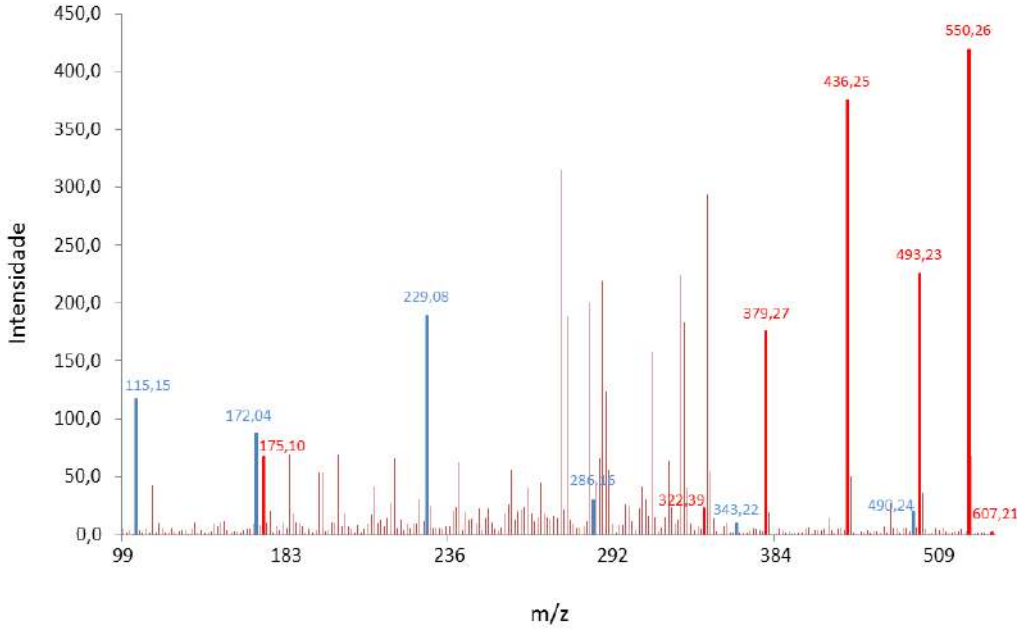


Figura 5: Espectro proveniente de um espectrômetro de massa; picos dos fragmentos tipo **b** e **y** em azul e vermelho, respectivamente

Alguns valores destacados em cores azuis e vermelhas salientam a *m/z* de determinados picos. Os íons apresentados no espectro em cores azuis e vermelhas representam fragmentos **b** e **y**, respectivamente. O que são os fragmentos do tipo **b** e **y** será explicado mais adiante. Os fragmentos **y** costumam ser mais intensos do que os fragmentos **b**. Um fato importante para se destacar é a quantidade de picos com

intensidades muito baixas, chamadas de ruídos. Esses ruídos podem ser provenientes de vários fatores, um deles é o processo de fragmentação em regiões aleatórias do peptídeo [12].

A primeira detecção de espectros *I* x *m/z* dos peptídeos de uma amostra inserida no espectrômetro é denominada MS^1, ou simplesmente MS. Os peptídeos que receberam carga (peptídeos ionizados) e foram atraídos para dentro do espectrômetro são inicialmente analisados e suas *m/z* e intensidades são registradas em espectros MS1. Assim, em espectros do tipo MS^1 serão encontradas todas as *m/z* detectadas na análise inicial das amostras de peptídeos. A Figura 6 mostra um exemplo de espectro MS^1, plotado em gráfico, com o registro das intensidades dos peptídeos ionizados no eixo das ordenadas e a *m/z* no eixo das abscissas.

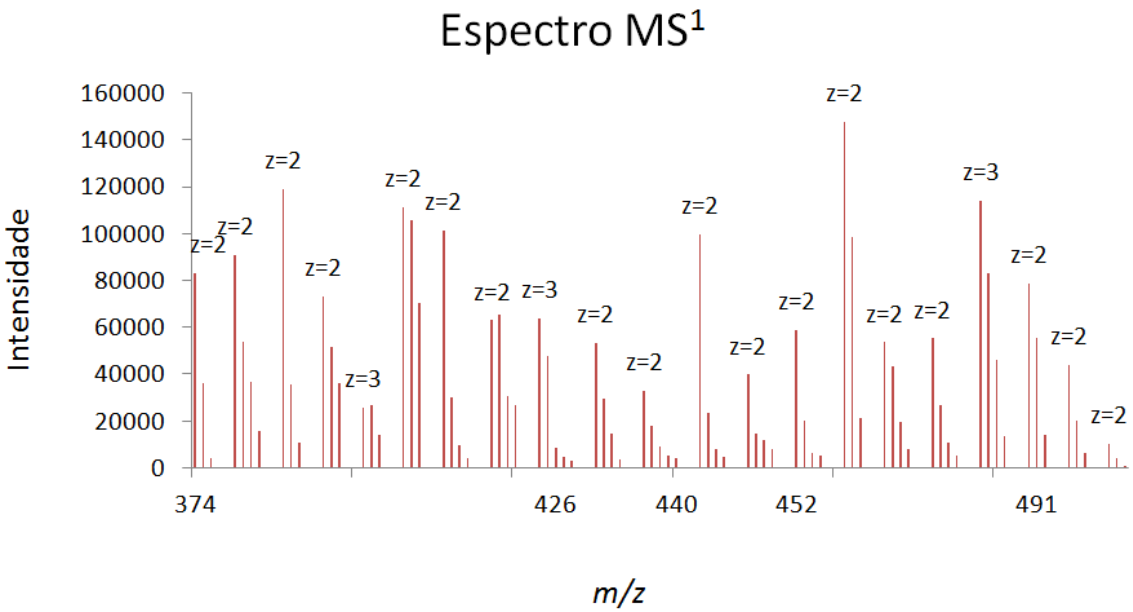


Figura 6: Espectro MS^1 proveniente de um espectrômetro de massa; *m/z* dos peptídeos, intensidades e suas respectivas cargas (*z*) sobre os picos monoisotópicos

Depois que este espectro é adquirido, o espectrômetro seleciona, de acordo com uma programação predeterminada pelo operador, um íon de determinada *m/z* de interesse, o precursor, e o fragmenta, registrando o espectro (*I* x *m/z*) dos íons. Estes espectros de íons-fragmentos de precursores são denominados MS^2. A unidade de massa dos peptídeos ionizados é o Dalton sobre carga (Da/z), denominado por alguns pesquisadores como Thomson (Th) [52]. Mas esta nomenclatura proposta ainda não é amplamente aceita pela comunidade de espectrometria de massa. Assim nos referiremos a esta unidade como "*m/z*".

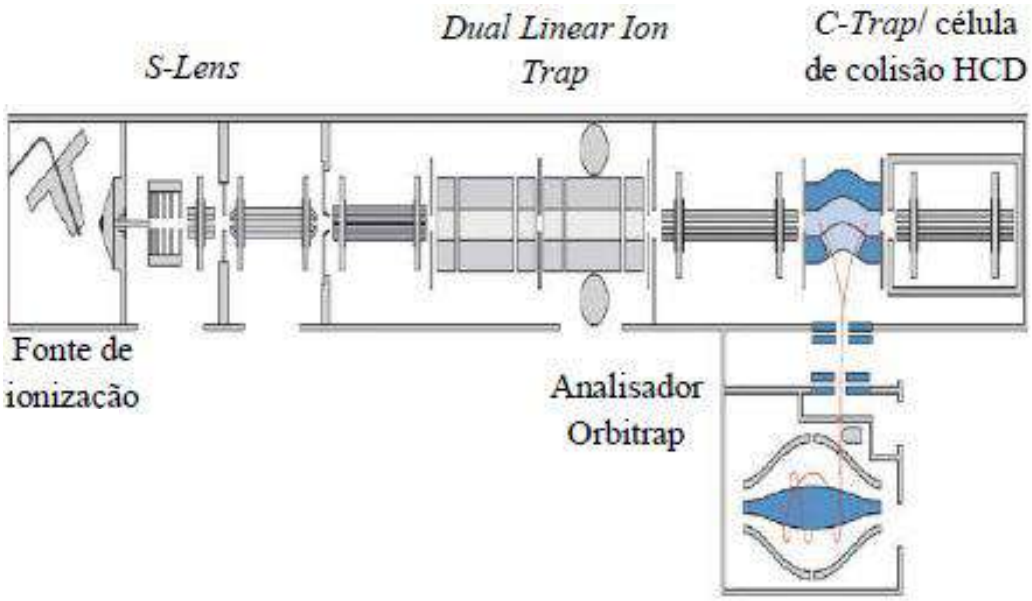


Figura 7: Desenho esquemático de um espectrômetro de massas *Linear trap* quadrupolo (LTQ) Orbitrap Velos [61], modificado por [62]

A Figura 7 esquematiza o espectrômetro usado neste trabalho, LTQ Orbitrap Velos (*Thermo Fisher Scientific*), que possui uma fonte de ionização ESI, dois analisadores: *Ion Trap Linear* e um Orbitrap, e os detectores, conforme mostrado no esquema. O *Ion Trap Linear* possui duas células lineares similares a do analisador quadrupolo que proporcionam um isolamento rápido e captura/seleção de íons para dissociação. As duas células têm a mesma fonte comum de RF e AC, mas tensões de corrente contínua independentes, permitindo transferência de íons de uma célula para outra. A primeira célula do *Ion Trap Linear* é operada em uma pressão maior por estar otimizada em aprisionar e dissociar íons. A segunda célula está otimizada para separação dos íons pelo *m/z* fornecidos pela primeira. O analisador Orbitrap aprisiona os íons para fragmentação. A presença desses dois tipos de analisadores no espectrômetro permite tirar vantagens de cada um. Geralmente obtém-se a massa dos precursores com alta resolução pelo Orbitrap e associando a alta velocidade do LTQ na dissociação e análise dos íons parentais [62].

A identificação de proteínas usando o espectro de massa [53] inclui o uso de espectro MS^2 dos fragmentos dos peptídeos precursores, da proteína, previamente detectados pelo espectrômetro. A identificação se dá pela comparação destes espectros com os espectros teóricos obtidos a partir das proteínas depositados em bancos de dados, empregando método que pontua estatisticamente as sequências mas próximas

daquele peptídeo. A utilização de fragmentos permitiu ampliar a identificação de proteínas presentes em um proteoma em uma só corrida cromatográfica e análise por espectrometria de massa.

Atualmente o uso do processo de fragmentação predomina na proteômica [14]. Para proteínas desconhecidas, no entanto, só há uma opção na espectrometria de massas para a sua identificação, o sequenciamento *de novo*. Este usa a diferença de massa entre os fragmentos para montar a sequência de aminoácidos do peptídeo, mesmo que parcialmente. A informação sobre a massa do peptídeo precursor, adquirido no espectro MS^1, e sobre as massas de seus fragmentos, adquiridos no espectro MS^2, é usada no sequenciamento *de novo*, como será mostrado a seguir.

Apesar de, na maioria dos casos, a identificação ser feita através de banco de dados, a evolução da técnica depende não só da evolução de espectrômetros de alta resolução, mais acurados, mas também da evolução de técnicas de interpretação de resultados *in silico*, o que permitiria a anotação de genes. O sequenciamento *de novo* é a motivação deste trabalho.

3.2 Espectros multiplex

Os espectros MS^2 são considerados multiplex quando contêm registrados íons-fragmentos de mais de um peptídeo. Estes espectros são complexos, pois contêm grande quantidade de picos *m/z*, o que dificulta a identificação dos peptídeos através de bancos de dados ou complica o sequenciamento *de novo*. A Figura 8 mostra um espectro multiplex onde foram registrados fragmentos de três peptídeos de carga z = 2.

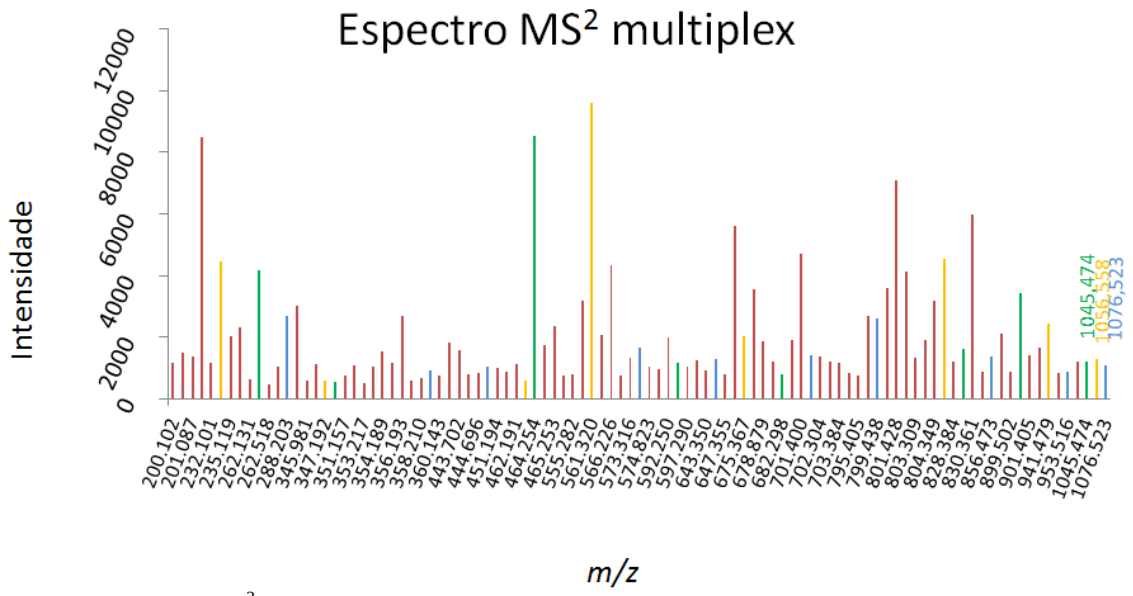


Figura 8: Espectro MS^2 multiplex contendo fragmentos de 3 peptídeos ionizados de carga z=2. Em verde os picos da série **y** do precursor de *m/z*=1045.474, em amarelo o precursor *m/z*=1056.558 e, em azul, precursor de *m/z*=1076.523

Espectros multiplex podem estar presentes em qualquer MS^2 para qualquer janela de fragmentação do MS^1. Mesmo quando um peptídeo precursor é isolado dentro de uma pequena janela de 2 Th no MS^1 para ser fragmentado, pode ocorrer que outro peptídeo, cuja *m/z* caia dentro da janela, e ele também será fragmentado nesta mesma janela. Há também técnicas de análise, como o DIA (*data independent analysis*), que usam janelas maiores de fragmentação aumentando a presença de íons-fragmentos de mais de um peptídeo no mesmo MS^2.

O grande desafio em se trabalhar com espectros multiplex reside em se lidar com maior volume de fragmentos, provenientes de diferentes precursores. Alguns fatores dificultam a interpretação dos espectros MS^2 multiplex, como por exemplo, a carga dos precursores analisados, a relação entre as intensidades dos peptídeos fragmentados, quando muito baixa pode ocorrer a supressão dos íons mais fracos, e, principalmente, a maior quantidade de íons-fragmentos presentes no espectro MS^2. Note que na Figura 8 podemos observar que os três picos mais intensos são de diferentes precursores, e também, que a quantidade de fragmentos do peptídeo representado na cor vermelha é muito maior do que a dos outros dois peptídeos, exemplificando outros fatores que dificultam a interpretação dos espectros MS^2.

3.3 Íons oriundos da fragmentação

Cada técnica de análise de peptídeos requer uma fragmentação específica que produz íons com características distintas. Em seu trabalho, CHI *et* al [33] afirmam que os dados oriundos do tipo de fragmentação HCD contêm mais informações sobre os fragmentos que o CID, incluindo o registro de picos de íons com *m/z* menores. Dessa forma, sem perda da generalização, a fragmentação por HCD será o foco deste trabalho por ser, atualmente, a mais usada, deixando alguns comentários sobre os fragmentos oriundos de outras técnicas de dissociação, como por exemplo, CID e ETD. A metodologia do sequenciamento *de novo* desenvolvida aqui serviria, a princípio, para qualquer forma de fragmentação.

Um peptídeo quando é fragmentado por HCD gera principalmente íons dos tipos **b** e **y** [15], como mostra a Figura 9. Estes íons provêm da quebra das ligações peptídicas, ligação entre o grupo carboxil de um resíduo de aminoácido com o grupo amino do resíduo do aminoácido seguinte feitas durante a tradução proteica.

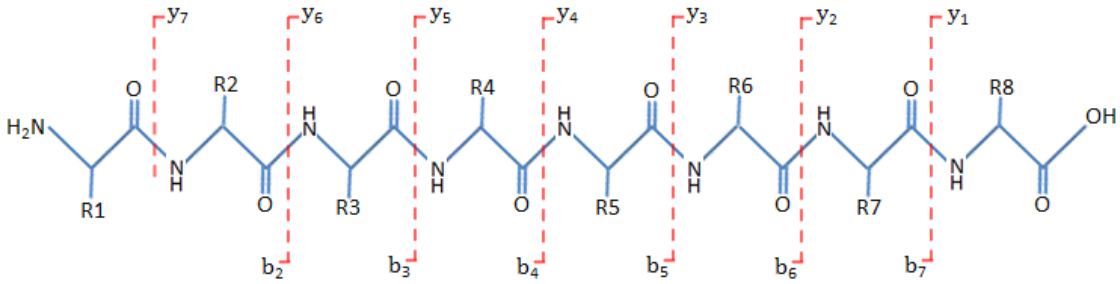


Figura 9: Esquema dos fragmentos tipo **b** e **y** de um peptídeo

Como exemplo, os íons b_2 e y_6 da Figura 9 surgem quando a colisão com o gás inerte cliva a segunda ligação peptídica, sendo b_2 o fragmento que se inicia no N-terminal e termina no C da 2ª ligação peptídica e y_6 o fragmento que inicia no N da 2ª ligação peptídica e termina no C-terminal do peptídeo; b_3,y_5 são os íons oriundos do corte da terceira ligação peptídica, e assim por diante. Genericamente, b_n, y_{N-n}, são os íons complementares originados do rompimento da n^{ésima} ligação peptídica, sendo *Nr* o número total de resíduos do peptídeo. Note que b_1 não é mostrado na Figura 9 por ser de difícil detecção por causa do envolvimento do C na clivagem e raramente aparece nos espectros.

Desta forma, o espectro MS^2, proveniente da amostra fragmentada de um determinado precursor, vai conter prioritariamente íons **b** e **y**, e mais alguns outros íons

de perdas neutras[5], imônios[6] e outros fragmentos da cadeia lateral. Na fragmentação por HCD, outros tipos de íons-fragmentos também podem ser registrados, como por exemplo, o íon tipo **a**, que será definido a seguir.

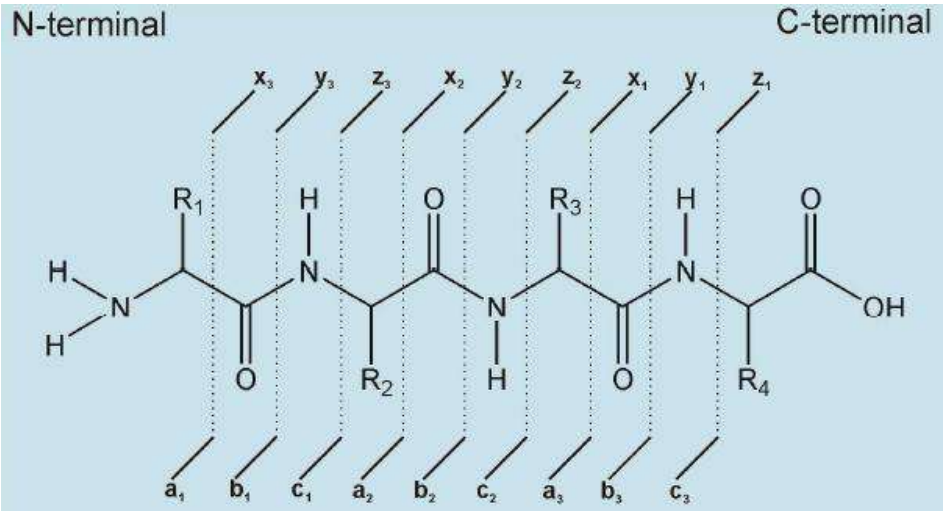


Figura 10: Fragmentos dos tipos **a**-**x**, **b**-**y** e **c**-**z** [16]

Apesar dos tipos de fragmentos **b** e **y** oriundos da fragmentação CID e HCD serem os mais comuns, dependendo do local da quebra efetuada no peptídeo, podem ser gerados pares de fragmentos dos tipos **a**-**x**, **b**-**y** e **c**-**z** [15], cujas clivagens acontecem em ligações químicas diferentes como ilustrados na Figura 10. Fragmentos **a**-**x** surgem da quebra das ligações covalentes entre o carbono alfa e o carbono da carboxila (Cα-C), principais íons gerados pelo EDD (e*lectron detachment dissociation*), fragmentos **c**-**z** são formados do corte após as ligações peptídicas, ligação entre o nitrogênio da amida e o carbono alfa (N-Cα), típicos de fragmentações por ETD e ECD. Um determinado fragmento pode não estar ionizado e assim não ter sua *m/z* registrada no espectro MS^2, o que irá representar uma das dificuldades do processo de sequenciamento *de novo*. Os tipos de íons **b** e **y** da fragmentação do precursor de carga +2 são mostrados na Figura 11**.**

[5]Perda neutra é uma eliminação de moléculas não carregada de um íon durante sua dissociação, por exemplo, perda de molécula de água e/ou amônia, H_2O (18Da) e/ou NH_3 (17Da) [56].
[6]Os imônios são íons com perda de molécula de CO na sua dissociação [57]. Eles são gerados quando a fragmentação ocorre em um único resíduo, cortando simultaneamente nas posições amino da ligação peptídica e antes da carboxi-terminal da próxima ligação peptídica, ligação Cα - C [58]

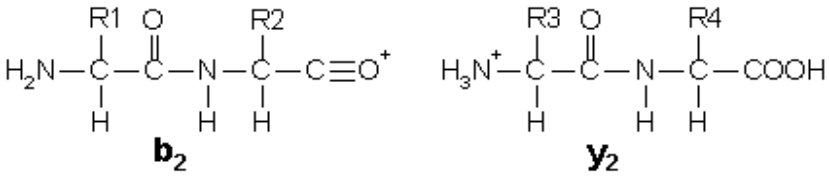


Figura 11: Íons dos tipos **b** e **y** gerados após a fragmentação de um precursor, z=+2 e N=4 [16]

Atendo-se somente aos íons **b**-**y**, e considerando um peptídeo precursor teórico monoisotópico de carga +2, cada íon fragmentado teria carga +1. A massa teórica destes íons será igual à soma das massas de resíduos de aminoácidos que o compõe, mais um hidrogênio no N-terminal no íon **b**, ou, considerando-se o íon tipo **y,** mais uma molécula de H_2O e mais uma carga H^+, como pode ser observado na Figura 11. Conforme Eq.(1) e (2), dadas abaixo, as massas teóricas dos íons-fragmentos de carga +1 do tipo **b** e **y** são:

$$m(b_i) = \sum_{j=1}^{i} m(r_j) + m(H) \quad (1)$$

$$m(y_i) = \sum_{j=Nr-i+1}^{Nr} m(r_j) + m(O) + 3m(H) \quad (2)$$

Onde $m(b_i)$ *e* $m(y_i)$ são as massas dos íons b_i e y_i, respectivamente, onde *i* é o índice do íon, *Nr* é o número total de resíduos do peptídeo precursor; $m(r_j)$ é a massa do j*ésimo* resíduo de aminoácido r_j do precursor, e *m(O)* e *m(H)* são as massas teóricas do oxigênio e hidrogênio, respectivamente.

Os termos que complementam a massa dos fragmentos a partir dos resíduos são chamados de *offset*. Nas Eq. (1) e (2) acima, os *offsets* são a massa do hidrogênio, $m(H)$ = 1.0073, e a soma das massas do hidrogênio e uma molécula de água, $m(O) + 3m(H)$ = 19.0183, respectivamente. A Tabela 2, criada por DANCIK [24], lista diversos tipos de íons provenientes da fragmentação CID, a frequência com que aparecem nos espectros e os respectivos *offsets*. Os resíduos ácidos perdem H_2O, enquanto os básicos perdem NH_3. Note que a massa do íon a_i é dada pela Eq.(3), e tem *offset* igual a $-\mathrm{m(H)} - \mathrm{m(O)} - \mathrm{m(C)} \cong -29.0027$.

$$m(a_i) = \sum_{j=1}^{i} m(r_j) + (-m(H) - m(O) - m(C)) \quad (3)$$

Tabela 2: Frequências e *offsets* de íons **a**, **b** e **y** obtidas por Dancik [24]; y2 representa um íon y com carga +2; os demais íons têm carga +1.

Offset	*Offset* inteiro	Frequência	C- ou N-terminal	Íon
18,85	19	0,6895	C	y
0,85	1	0,6484	N	b
-17,05	-17	0,3858	N	b-H_2O
0,9	1	0,28,31	C	y- H_2O
-27,15	-27	0,2329	N	a
20,05	20	0,2089	C	y2
-16,15	-16	0,1815	N	b-NH_3
1,9	2	0,1495	C	y-NH_3
-35,2	-35	0,1724	N	b-H_2O-H_2O
-34,2	-34	0,153	N	b-H_2O-NH_3
-44,25	-44	0,1473	N	a-NH_3
-45,15	-45	0,1221	N	a-H_2O
2,3	2	0,1164	C	y2-H_2O
-16,1	-16	0,1107	C	y-H_2O-NH_3
-17,15	-17	0,1039	C	y-H_2O-H_2O

As massas teóricas de dois íons complementares de carga +1 cuja fragmentação ocorreu na $k^{\text{ésima}}$ ligação peptídica, são dadas por:

$$m(b_k) = \sum_{j=1}^{k} m(r_j) + m(H) \tag{4}$$

$$m(y_{Nr-k}) = \sum_{k+1}^{Nr} m(r_j) + m(O) + 3m(H) \tag{5}$$

A soma das Eq. (4) e (5) compõe a massa do precursor de carga +2, $m(P^{+2})$. A *m/z* teórica do peptídeo precursor de carga +2 pode ser expressa como:

$$\frac{m(P^{+2})}{2} = \sum_{i=1}^{Nr} \frac{m(r_i)}{2} + \frac{4m(H)+m(O)}{2} = \frac{m(b_k)+m(y_{Nr-k})}{2} \tag{6}$$

Genericamente, a uma equação para $m(P^{+z})/z$ poderia ser escrita para qualquer carga +z do precursor.

3.4 Cálculo dos resíduos da sequência

O espectro de massa experimental registra *m/z* próximo da massa teórica da Eq. (6) dependendo da resolução e da acurácia do espectrômetro, o que vai exigir a adoção

de um valor de tolerância máximo para se aceitar um pico como o registro de um provável fragmento.

É usual trabalhar na formulação teórica diretamente com as massas dos fragmentos e precursores, ao invés de massa-carga, e ainda padronizar as cargas, +1 para fragmentos e +2 para precursores, no sequenciamento *de novo* de peptídeos. As massas-carga experimentais dos íons fragmentos e dos precursores são ajustadas quando necessário para serem introduzidas nas expressões teóricas. Assim, se algum íon da série **b** ou **y**, agora com carga +1, for identificado no espectro MS^2, seu complemento pode ser calculado diretamente pela relação definida na Eq.(7), seja qual for a massa do precursor, como reproduzido abaixo.

$$m(y_{Nr-k}) = m(P^{+2}) - m(b_k) \tag{7}$$

Vale observar que todo peptídeo precursor, por ser um íon, possui pelo menos uma carga. Quando essa carga é +1 no momento da fragmentação, apenas o fragmento ionizado é registrado perdendo-se assim todos os fragmentos de uma das séries, tipo **b** ou **y** (protonado no alfa amino ou epsilon amino da lisina). Por este motivo, os operadores tendem a ajustar o espectrômetro para fragmentar peptídeos precursores de carga maior ou igual a +2 (protona no alfa amino e/ou no epsilon amino da lisina ou grupo guanidina da arginina).

Generalizando, a massa de um resíduo pode ser expressa tanto pelas diferenças entre fragmentos consecutivos do tipo **b**, quanto do tipo **y**, da seguinte forma:

$$m(r_i) = m(b_i) - m(b_{i-1}) \tag{8}$$

$$m(r_{Nr-i+1}) = m(y_i) - m(y_{i-1}) \tag{9}$$

A massa do mesmo resíduo $m(r_i)$, expressa pela Eq.(8), pode ser calculada através da série **y** pela expressão:

$$m(r_i) = m(y_{Nr-i+1}) - m(y_{Nr-i}) \tag{10}$$

Note que se *i=1*, as Eq.(8) e (9) teriam que lidar com valores $m(b_0)$ e $m(y_0)$, que poderiam ser inferidos como *offsets* de **b** e **y**, respectivamente. O mesmo ocorreria se *i=Nr*, $m(b_{Nr})$ e $m(y_{Nr})$, e, neste caso, poderíamos teoricamente definir os valores como,

$$m(b_{Nr}) = m(P^{+}) - offset(\boldsymbol{y}) \quad (11)$$

$$m(y_{Nr}) = m(P^{+}) - offset(\boldsymbol{b}) \quad (12)$$

Entre os espectrometristas de massa é dito que, em um loop de sequenciamento *de novo* de *i=0,Nr*, a diferença entre picos de massa à esquerda da Eq.(8) sequenciaria o peptídeo corretamente, da esquerda para a direita do eixo ordenado do espectro, enquanto a diferença da Eq.(9) sequenciaria da direita para a esquerda, sequenciando o peptídeo de trás para frente. Ambos os processos, no fundo, são idênticos e representados pela Eq. (9) e (10).

Assim sendo, os fragmentos registrados no MS^2 permitem que se reconstrua a correta sequência de aminoácidos do peptídeo que deu origem a esses fragmentos, peptídeo precursor, se todos os íons fragmentados de pelo menos uma das séries **b** ou **y** estiverem presentes no espectro. Existem fatores que dificultam a reconstrução da sequência correta, como será abordado adiante no capítulo 4.

4 SEQUENCIAMENTO *DE NOVO* DE PEPTÍDEOS

Reconstruir a correta sequência de aminoácidos de um peptídeo a partir de seus fragmentos registrados no MS^2, conhecida como sequenciamento *de novo*, fundamenta-se nas diferenças de *m/z* entre os fragmentos do íon precursor registrados nos espectros, como ilustrado na Figura 12.

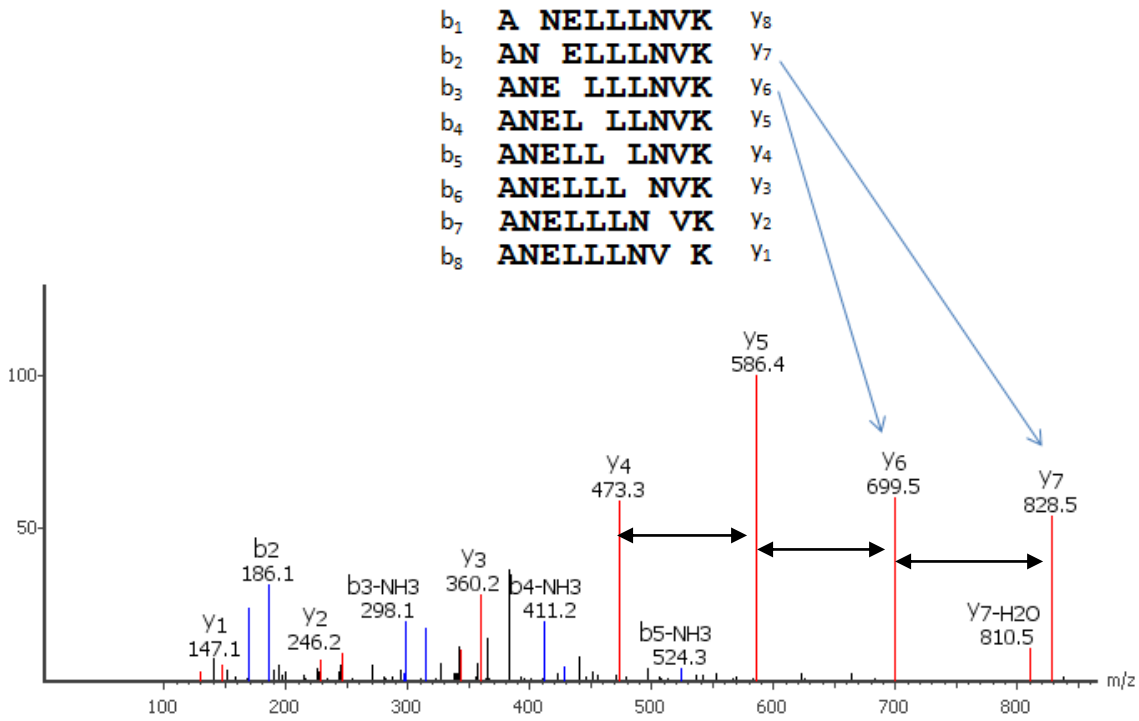


Figura 12: Espectro MS^2 do peptídeo ANELLLNVK, onde alguns íons das séries dos fragmentos **b** e **y** estão anotadas[25]

A Figura 12 mostra um espectro onde os íons-fragmentos do tipo **b** e **y** de um peptídeo precursor estão anotados. As massas desses íons são complementares até a massa do precursor, ou seja, somando-se a massa de um íon b_i e a massa do íon complementar y_{N-i} temos a massa do precursor, com uma variação muito pequena para a massa teórica, como já foi formalizado na Eq. (6). Os picos dos íons do tipo **y** são, com frequência, mais intensos que os **b** nos espectros MS^2, ou seja, mais abundantes. O pico de b_1 raramente está presente e o íon tipo **a** é muito menos frequente e quando aparece tem pico de menor intensidade.

Outros tipos de fragmentos também podem estar presentes no espectro, como as massas de fragmentos com perdas neutras, listados na Tabela 2. Considerando que a *m/z* de um íon genérico de carga +z seja

$$\frac{m(ion^{+})}{z} = \frac{1}{z}\sum m(r_j) + \frac{\Delta}{z} \quad , \tag{13}$$

o acréscimo de massa-carga, Δ/z a ser adicionada à soma de massas dos resíduos, $\sum m(r_j)\,/z$, que compõe o tipo de íon em questão, é o valor *offset/z*.

Tabela 3: Tipos de íon considerados neste trabalho, incluindo os que apresentam perdas neutras, na fragmentação por HCD e os respectivos *offsets*. O número e/ou sinal sobreescito no "Tipo de íon" representa a carga

Tipo de íon	*Offset* massa-carga, Δ/z
y^+	19,02
b^+	1,01
a^+	-26,98
Y^+-NH_3	1,99
Y^{+2}	10,01
Y^+-H_2O	1,01
b^+-H_2O	-16,99
b^+-NH_3	-16,01

Para se chegar, por exemplo, ao valor *offset* de Y^+-NH_3, subtrai-se a massa da perda da molécula NH_3, que é 17,0265, da massa do *offset* do tipo y^+, que é dada pela massa da molécula OH_3, 19,0183. Assim, o *offset* para o tipo Y^+-NH_3 é dado por 19,0183-17,0265=1,9918. O mesmo procedimento ocorre para os outros *offset*, sempre levando em consideração o *offset* do tipo b ou y.

Os *offsets/z* usados neste trabalho são listados na Tabela 3. Importante aqui definir que cada pico analisado do espectro MS^2 tem uma série de evidências que imputam a ele uma maior probabilidade de ser um pico do tipo **y** ou **b** do peptídeo procurado. Se existe um outro pico que somado ao pico analisado complementa a massa do precursor, este fato é uma forte evidência de que o pico analisado seria parte da sequência do precursor, seja **b** ou **y**. Se há um pico que possa ser um íon **a**, relacionado ao pico analisado, este fato também constitui uma evidência da existência deste pico analisado como parte da sequência procurada. Um pico referente a uma perda neutra de H_2O e NH_2 relacionada ao pico analisado também pode ser considerada evidência. Assim como é evidência a existência de um imônio reforçando que um aminoácido encontrado pela diferença entre dois picos do espectro MS^2 faça mesmo parte da sequência procurada. Os íons imônios [28], se presentes nos espectros aumentam a probabilidade de confirmar a presença dos resíduos correspondentes na sequência. Ou

seja, quando a fenilalanina está presente no peptídeo, não raro seu imônio aparece no espectro, confirmando sua presença. A Tabela 3 mostra os íons que podem estar presentes em um MS^2 oriundo de uma fragmentação por CID ou HCD, todos podendo ser evidências da pertinência de um fragmento na sequência procurada. A Tabela 4 lista a massa dos resíduos teóricos (sem carga) e dos seus imônios (carga +1) correspondentes, utilizados neste trabalho.

A carga do precursor é inferida pela diferença entre as *m/z* dos picos, no espectro de MS^1, do envelope isotópico do mesmo peptídeo. Se a carga do precursor for +2, a diferença entre os íons-peptídeos isótopos será de 0,5, já que o registro *m/z* de um isótopo com um carbono 13 será igual à massa do peptídeo monoisotópico mais 1 Da dividido por 2, referente à carga. A carga é determinada então pelo inverso das diferenças constantes entre os picos do envelope isotópico do espectro de MS^1. Assim pode-se obter a *m/z* monoisotópica do peptídeo precursor, o primeiro pico da série de picos que compõem o envelope isotópico, e sua carga z, calculada automaticamente pelos programas acoplados ao espectrômetro que ajudam na análise e na aquisição dos espectros.

Tabela 4: Lista de aminoácidos, massas teóricas monoisotópicas dos resíduos e massa de seus imônios associados [59]

Aminoácido	Símbolo	Massa monoisotópica teórica do resíduo, Da	Massa do Imônio de carga +1, Da
Glicina	G	57,02146	30
Alanina	A	71,03711	44
Serina	S	87,03203	60
Prolina	P	97,05276	70
Lalina	V	99,06841	72
Treonina	T	101,14768	74
Cisteína	C	103,00919	76
Leucina	L	113,08406	86
Isoleucina	I	113,08406	86
Aspargina	N	114,04293	87
Ácido aspártico	D	115,02694	88
Glutamina	Q	128,05858	101
Lisina	K	128,09496	101
Ácido glutâmico	E	129,04259	102
Metionina	M	131,04048	104
Histidina	H	137,05891	110
Fenilalanina	F	147,06841	120
Arginina	R	156,10111	129
Tirosina	Y	163,06333	136
Triptofano	W	186,07931	159

A técnica de sequenciamento *de novo* não é nova, mas permanece ainda como um grande desafio para a proteômica, especialmente quando se pensa em organismos com genoma não sequenciado ou não anotado [37].

A relevância de novas técnicas para sequenciamento *de novo* está na possível identificação de proteínas cujas sequências não estejam depositadas em bancos de dados, e também contribuir na anotação de proteínas de um genoma [37].

Neste trabalho serão utilizadas, então, duas informações básicas para fazer o sequenciamento *de novo*, i.e., a massa peptídeo precursor, obtida no espectro MS^1, e o espectro de massa dos fragmentos deste precursor, MS^2.

O sequenciamento *de novo* parece simples a princípio, mas esbarra em alguns problemas. São eles:

- nem todas as massas *m/z* dos fragmentos podem estar presentes no espectro ou ter uma intensidade baixa demais, confundindo-se com ruídos;
- não se sabe previamente qual fragmento é de que tipo, se **b** ou **y**;

- a diferença entre fragmentos não é exatamente a massa teórica do resíduo; por esse motivo há sempre um valor de tolerância δ_{max} pré-definido a ser considerado;
- há muito ruído nos espectros, picos cujas diferenças de massas indicam uma massa teórica de resíduos dentro da tolerância δ_{max} adotada, mas que não representa um resíduo do peptídeo procurado;
- considerando todas as diferenças de massa dos picos do MS^2 que representam um resíduo, muitas combinações de sequências de aminoácidos, que somam a massa total do precursor, são encontradas.

Decidir qual a sequência mais provável é um grande desafio. O espaço de busca do problema pode ser muito grande. Considerando que este problema seja representado por um grafo, para se avaliar a melhor sequência é necessário adotar uma função de pontuação de cada resíduo de aminoácido das sequências candidatas neste grafo. O valor máximo da soma destas pontuações definiria a melhor sequência, supostamente o peptídeo procurado, direcionando a escolha. A eficiência da busca depende de uma avaliação que represente o problema corretamente para garantir o sucesso do sequenciamento *de novo*. A escolha da função de pontuação do problema não é tarefa trivial e será abordada mais adiante.

Para instrumento do tipo *IonTrap*, um fator que dificulta o processo do sequenciamento *de novo* é a regra do 1/3 da massa do precursor [29], que faz com que algumas massas iniciais (baixas) não sejam registradas. Isso se deve à limitação do equipamento para a determinação de íons com valores de *m/z* abaixo de 1/3 da *m/z* do precursor. O valor da voltagem inicial da rádio-frequência para estabilizar um íon impõe que a sua *m/z* não seja muito baixa. Quanto mais próximo da fronteira de instabilidade estiver um íon, menor será a chance de ele ser registrado [29].

4.1 Heurísticas para sequenciamento *de novo*

Programas para sequenciamento *de novo* usando diferentes abordagens têm sido apresentados ao longo das últimas três décadas. Esses programas utilizam os mais variados formatos de arquivos de entrada, onde os predominantes são MGF e MS^2 [38], todos construídos a partir dos dados RAW fornecidos pelos espectrômetros.

Alguns trabalhos, os mais expressivos de uma lista de muitos programas, serão brevemente descritos neste item. O primeiro programa para sequenciamento *de novo* de peptídeos, chamado de PAAS3, usava a série **y** e encontrava as combinações teóricas dos aminoácidos a partir da massa do precursor [18]. Inicialmente eram geradas as possíveis combinações usando os 20 aminoácidos que resultavam na massa do precursor. Uma função retornava a sequência combinada que melhor representava o peptídeo conforme presença dos picos no espectro experimental, usando a série C-terminal.

O sequenciamento *de novo* era uma tarefa demorada devido o grande poder computacional demandado e limitação dos processadores. Com o objetivo de ajudar na identificação de peptídeos em banco de dados de proteínas, surgiu um outro tipo de abordagem conhecida como *sequence tags*, que é o sequenciamento *de novo* de pequenas sequências contínuas de aminoácidos. Usando a mesma estratégia de combinação de aminoácidos e posterior comparação com os picos do espectro, apresentada por SAKURAI [18], ISHIKAWA [20] usou apenas os picos mais intensos do espectro MS^2 para determinar as *tags*. A grande diferença é que ele considerou tanto as séries N-terminais quanto as C-terminais. TABB [32] e LI [34] foram os primeiros a usarem grafo para encontrar *sequence tag* para ajudar na identificação de possíveis proteínas em bases de dados de proteínas catalogadas.

Utilizar a teoria dos grafos para representar as possíveis sequências *de novo* para um peptídeo de um determinado espectro MS^2 foi introduzido por BARTELS [22], onde cada pico do espectro corresponde a um nó do grafo e as arestas apenas existem quando a diferença de massas entre dois picos resultassem na massa de um resíduo de aminoácido conhecido. No trabalho de BARTELS, os caminhos encontrados, tanto para as séries C-terminais quanto N-terminais, eram avaliados de acordo com a probabilidade obtida a partir de dados de treinamento para os diferentes tipos de fragmentações. A introdução do grafo direcionado [36] como representação fez com que o sequenciamento *de novo* de peptídeos fosse visto como um esforço para encontrar o melhor caminho em um grafo da menor para a maior massa. O programa Lutefisk [23], um programa para identificação de peptídeos em banco de dados, usou o sequenciamento *de novo* de peptídeos para ajudar na identificação.

O uso de grafo para tentar resolver o problema de sequenciamento *de novo* e considerar as séries N-terminal e C-terminal norteou as soluções do sequenciamento *de*

novo, como o programa SHERENGA, apresentado por DANCIK [24]. As bases das soluções para o problema de sequenciamento *de novo* de peptídeos não divergem muito das apresentadas no programa SHERENGA. Ele apresenta uma lista de evidências a partir das massas dos fragmentos de um espectro. Ainda hoje os principais programas usam grafo e utilizam-se das evidências descritas por DANCIK [24]. As evidências são informações que dão suporte na descoberta de um pico. Por exemplo, suponha um determinado pico de massa inteira 175 e um peptídeo de massa 664. Existirá um pico complementar de massa 490 e, sendo este pico um fragmento do tipo **b**, poderá existir um pico de massa que represente a perda de uma água e um pico de massa do tipo de fragmento **a**, com números inteiros 472 e 462, respectivamente. Esses dois picos seriam evidências de que o pico de massa 490 é um fragmento do tipo b.

Uma abordagem para sequenciamento *de novo* utilizando Programação Dinâmica foi apresentada por CHEN *et* al [26] para a busca do melhor caminho. Essa busca era baseada nas intensidades dos picos e nas evidências dos picos do espectro MS^2 na tentativa de encontrar uma solução ideal. O programa PepNovo, apresentado por FRANK [27], também utiliza grafo usando algumas das evidências elencadas por DANCIK [24], e uma busca dos melhores caminhos usando uma Programação Dinâmica. Este programa foi usado durante um momento do desenvolvimento deste trabalho, mas os resultados não foram bons, e por isso o programa foi suprimido na obtenção dos resultados finais.

Recentemente, JEONG [35] apresentou o programa UniNovo. Essa abordagem utiliza um grafo tradicionalmente conhecido para o sequenciamento *de novo* de peptídeos e uma busca usando Programação Dinâmica. Adota também as probabilidades resultantes de dados de treinamento prévio que registra os pesos das evidências dos picos presentes no espectro se cada pico for considerado como um tipo de fragmento diferente. Juntamente com um estes pesos, adiciona-se um sistema de pontuação para os nós baseado também nas intensidades dos picos do espectro. Diferentemente da pontuação proposta por DANCIK, a pontuação do programa UniNovo acrescenta o logaritmo da razão da probabilidade do pico ser verdadeiro e do pico ser falso. Este programa necessita que seja informada a quantidade de aminoácidos da sequência que se deseja considerar para determinar os caminhos candidatos. Por esse motivo não se utilizou esse programa nas comparações realizadas neste trabalho.

Inúmeras abordagens foram apresentadas para resolver o problema de sequenciamento *de novo* [22], cada heurística apresentando um desempenho diferente [41]. Avanços da espectrometria na aquisição de espectros de melhor qualidade e o aumento da capacidade de processamento do hardware fez com que o sequenciamento *de novo* emergisse novamente para o centro das atenções [42].

JEONG [35] aponta o programa comercial apresentado por MA [25], Peaks, como algoritmo do estado da arte. Além deste, usaremos neste trabalho o programa pNovo [33] e o Novor [49] recentemente lançado, que utiliza árvore de decisão para realizar o sequenciamento *de novo* de peptídeo. Estes programas serão utilizados para comparação do programa DNbuilder, desenvolvido neste trabalho. Uma breve explicação dos programas será apresentada a seguir, onde as notações seguem as dos artigos.

4.1.1 Peaks

O programa comercial Peaks, apresentado por MA *et* al [25], realiza um filtro de picos considerados ruídos, mas não detalha essa etapa. Considerando as intensidades dos tipos de íons fragmentos x, y, y-H_2O, y-NH_3 ou a, b, c, b-H_2O, b-NH_3, o Peaks constrói as sequências candidatas a partir de todas as combinações possíveis entre picos cujas diferenças entre suas massas resultem na massa teórica de um aminoácido conhecido, pontuando todas as sequências usando suas intensidades para selecionar inicialmente as dez mil melhores combinações de sequências. Depois refina a pontuação final de acordo com uma função de bonificação, cujo valor obtido de bonificação é somado à pontuação inicial da sequência candidata.

Para obter o valor da bonificação das sequências, inicialmente calcula-se a abundância relativa de cada pico y ou b considerando os tipos de fragmentações das evidências x, y-H_2O, y-NH_3 ou a, c, b-H_2O, b-NH_3. Para se chegar a abundância relativa de um pico evidência, divide-se a intensidade do pico evidência pela intensidade do pico observado **y** ou **b**. Também é usado na bonificação o erro quadrático de acurácia baseado na tolerância usada no cálculo das massas dos fragmentos, que é um fator calculado dividindo-se o erro calculado pela tolerância δ_{max}. O logaritmo da abundância do pico é a última variável da bonificação. A função de bonificação da pontuação de cada pico é mostrada na Eq. (14).

$$bonificação = f\left(\frac{h1}{h}\right) * f\left(\frac{h2}{h}\right) * f\left(\frac{h3}{h}\right) * exp\left(-\left(\frac{m^{'}-m}{\delta}\right)^{2}\right) * \log h \quad (14)$$

Onde: h é a intensidade do tipo de fragmento *b/y*,

h_x representa a intensidade de cada tipo de evidência do fragmento **y**/**b** presente no espectro. Para a série **y** os tipos considerados são x, y-H_2O e y-NH_3; Para a série **b** os tipos considerados são a, c, b-H_2O e b-NH_3, tendo, portanto mais um termo adicionado à bonificação de pontuação, $f\left(\frac{h4}{h}\right)$. Tomando-se um pico como base, verifica-se a maior ocorrência de picos no espectro que coincidam com o *offset* das evidências dos fragmentos considerados para se saber se pertence à série **b** ou **y**.

$f\left(\frac{h_x}{h}\right)$ é a função para determinar a intensidade relativa de um pico evidência do espectro, calculada somente se o pico estiver presente;

$(m^{'} - m)$ representa as massas calculada e teórica do fragmento do íon sob questão, respectivamente, e

δ é a massa limite de tolerância adotada: max admissível para $(m^{'} - m)$.

A bonificação da Eq. (14), calculada para cada pico, é somada à intensidade do referido pico. Assim, com o refinamento da pontuação dos picos das sequências através da bonificação, as sequências candidatas são reordenadas dentro da lista.

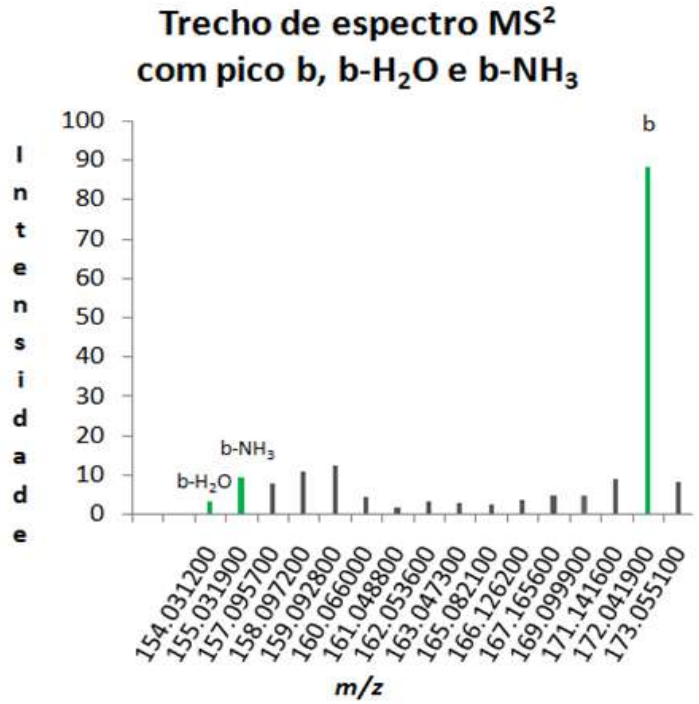


Figura 13: Gráfico dos picos de um espectro que participam da pontuação final do pico de *m/z* 172,0419 para o programa Peaks

A Figura 13 mostra uma janela do espectro de massa MS^2 com os picos que serão usados na pontuação de um suposto pico *b* de uma determinada sequência investigada, sequência montada com possíveis fragmentos tipo **b**. Nesse espectro são apresentados os picos que participam da equação para cálculo da pontuação final do pico de massa 172,0419, usando somente a série **b**. O pico *b* da sequência candidata é o de *m/z* 172,0419 e sua intensidade de 88, o *h* da Eq.(14). Evidências da veracidade deste pico **b**, os dois picos encontrados no espectro e que participam da pontuação, são 154,0312 e 155,0319, íons do tipo **b**-H2O e **b**-NH_3, respectivamente. Os rótulos em verde identificam, na Figura 13, estes íons. A intensidade do pico 154,0312 é 3 e a do pico 155,0319 é 9,2. Esses picos participam da pontuação do pico referente ao íon fragmento de tipo **b**. A bonificação retornada pela função para o referido pico é 0,06512, resultado da expressão temos (3/88*9,2/88*1) *exp(-((57,02146-56,8967/0,5)^2) * log(88).

A pontuação final de cada sequência candidata é calculada através do somatório da pontuação bonificada de cada pico da sequência candidata. A melhor sequência candidata é a que maximiza a pontuação.

4.1.2 pNovo

O programa pNovo, apresentado por CHI *et* al [33], realiza um pré-processamento nos espectros, onde os picos imônios são retirados. Depois disso, somente são aproveitados para o restante do processamento os 150 picos mais intensos, como *default*, ou um número *K* de picos mais intensos se o usuário preferir definir o parâmetro. Os picos são, então, ordenados de acordo com suas intensidades, onde o mais intenso estará na posição relativa 1 da lista, e o menos intenso na posição relativa final.

O programa cria um grafo direcionado usando os picos da série **b**, onde os picos aproveitados do espectro se tornam nós do grafo. Usando seis diferentes tipos de evidências de íons, a saber, y, b, y-NH_3, y-H_2O, a e y2, quando houver a presença de pelo menos duas destas evidências para um determinado pico selecionado (supondo que seja tipo **b)**, as intensidades dos picos das evidencias são somados às intensidades do pico selecionado, consolidando as intensidades numa só *m/z*.

Em seguida é feita a busca usando o algoritmo DFS [43], que gera uma lista de sequências candidatas, onde a pontuação inicial da sequência, chamada de *melhor-pontuação*, é dada pelo somatório das intensidades dos picos do caminho. A pontuação final de cada sequência candidata S do peptídeo P considera três diferentes premissas: maior-intensidade (S_H), local-fragmentado (S_F) e desvio-de-massa (S_{MD}).

O valor da pontuação de S_H refere-se aos picos mais importantes, onde o mais intenso recebe o valor de relatividade 1 e o menos intenso recebe o valor de relatividade *K*. Quanto menor for o valor da relatividade de um pico, maior será sua intensidade, consequentemente, maior será a pontuação calculada. Uma unidade é acrescida para cada evidência encontrada no espectro a partir de um pico base. O pNovo usa 6 evidências, i=6. Assim, considerando um pico p_i da sequência candidata, verifica-se cada acerto de acordo com a presença da sua evidência p_j observada: $match(pj) = 1, se\ "true"\ e\ zero\ "otherwise"$. A função *match* retorna para cada pico p_i a existência ou não de uma evidência p_j. Um pico que tiver todos os 6 picos evidências presentes no espectro fará com que a função *match* retorne 1 para cada um desses 6 picos. A pontuação S_H, para um peptídeo *P* de um espectro *S*, é dada pela equação:

$$S_H(S,P) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{i}\sum_{j=1}^{i} match(pj) \qquad (15)$$

Já a pontuação S_F baseia-se na quantidade total de clivagens de *P* observadas em *S*, *cf*, no maior *tag* em número de aminoácidos do peptídeo P, *tf*, e na quantidade de resíduos de aminoácidos da sequência candidata *P* observada, *length(P)*. A pontuação S_F é, portanto, dada pela expressão

$$S_F(S,P) = \frac{\sqrt{cf * tf}}{length(P)-1} \tag{16}$$

Para se chegar à pontuação S_{MD} referente às *k* maiores intensidades utiliza-se a equação

$$S_{MD}(S,P) = (T - \sqrt{\frac{\sum_{k\ maiores\quad intensidades\quad de\ p} md^2(p)}{K}})/\mathrm{T} \tag{17}$$

onde: *T* é a tolerância máxima admitida para a diferença entre o aminoácido teórico e o calculado pela diferença de dois picos,

md é o erro entre a massa do pico *p* observado e o íon teórico correspondente,

p é um dos picos de maior intensidade avaliado, e

K é uma quantidade pré-definida pelo usuário de picos mais intensos de *P*

Depois de calculadas as pontuações S_H, S_F e S_{MD}, a pontuação final de cada candidato *P* é dada pela equação

$$CScore(S,P) = (\sqrt[3]{S_H(S,P) x S_F(S,P) x S_{MD}(S,P)} x S_0 \tag{18}$$

onde S_O é o somatório das intensidades dos picos da sequência normalizada pela maior pontuação inicial, *melhor-pontuação*.

4.1.3 Novor

O programa Novor foi recentemente apresentado por MA [49] e apresenta uma abordagem diferente das usuais. O Novor não usa grafo e é baseado em árvores de decisão a partir de probabilidades retiradas de espectros de treinamento usando os íons da série **b**. Um conjunto de dados de treinamento é usado para determinar probabilidades baseadas em certas características dos picos do espectro. As probabilidades treinadas estão embutidas no código do programa. O Novor utiliza um

sistema de pontuação chamado de *feature-score*, que usa as características observadas de cada pico, onde função *feature-score* retorna uma pontuação. As características são:

- Intensidade do pico observado dividido pelo pico mais intenso;
- Quantidade de pico com intensidade maior ou igual à intensidade do pico observado;
- Quantidade de pico com intensidade maior ou igual à metade da intensidade do pico observado;
- Quantidade de pico com intensidade maior ou igual à intensidade do pico observado dentro de uma janela de 50Da;
- Quantidade de pico com intensidade maior ou igual à metade da intensidade do pico observado dentro de uma janela de 50Da;
- Intensidade do pico observado dividido pelo pico mais intenso dentro de uma janela de 50Da.

A pontuação é dividida em duas partes: pontuação de fragmentação e pontuação de resíduo. A pontuação de fragmentação resulta em valores diferentes de acordo com a evolução das perguntas respondidas. Perguntas diferentes fazem caminhar em ramos diferentes nas árvores de decisão com valores de probabilidades calculadas. Essas mesmas perguntas podem possuir probabilidades diferentes em lugares diferentes das árvores de decisão. As probabilidades dependerão do conjunto de perguntas e respostas ao longo da árvore.

A pontuação do resíduo baseia-se na probabilidade de um resíduo ser verdadeiro para o caminho. Num total de oito características diferentes considerando nove diferentes evidências resultam em 72 características para cada série, a saber, série **b** e série **y**, totalizando 144 características. Além dessas, a massa do precursor, a carga, a massa do íon y e a massa do íon **b**, são usadas como características.

Não é possível determinar como se chega à pontuação de cada pico sem os valores das probabilidades *feature-score* geradas pelo programa Novor por se tratar de código interno.

Diferentemente dos outros programas, que apresentam uma lista contendo as sequências candidatas ordenadas da maior pontuação para a menor, o Novor apresenta

uma única sequência candidata para cada espectro usando as técnicas de programação dinâmica.

4.2 Descomplexação de espectros multiplex

Para que espectros multiplex sejam registrados, dois ou mais peptídeos devem estar presentes em uma mesma janela de aquisição. O tamanho da janela depende de parametrização do espectrômetro de massa. Acontece que os íons registrados nos espectros MS^1 para um mesmo peptídeo podem ter cargas diferentes. Isso cria uma dificuldade a mais em analisar espectros multiplex, tendo que lidar com fragmentos de carga diferentes no mesmo espectro MS^2. KRYUCHKOV *et* al [2] apresentaram uma metodologia que, após deconvoluir e deisotopar os íons b, c, y e z dos espectros MS^2, combina pares de íons fragmentos para determinar os possíveis pares complementares dos íons cuja *m/z* caia dentro da janela de fragmentação ± uma margem de tolerância. Numa aquisição DDA em janelas de 2,5 *m/z*, os autores acharam outros peptídeos que não o precursor, co-eluídos na mesma janela de isolamento centrada no precursor. Esse experimento foi aplicado em tipo de fragmentações CID-HCD, e CID-HCD/ETD. Os pares complementares usados foram y-b (CID e HCD) e c-z (ETD).

Para facilitar a identificação dos peptídeos encontrados na janela, designados alvos, os espectros MS^2 são modificados, isto é, descomplexados, gerando-se um espectro para cada peptídeo alvo. O espectro gerado para cada alvo é construído aumentando-se as intensidades dos picos complementares do alvo, de acordo com uma classificação prévia dos picos complementares do MS^2 em três intervalos de confiança, I, II, III. O cálculo do pertencimento dos complementos a um intervalo de confiança é feito através de uma tabela construída a partir da observação de um conjunto expressivo de espectros MS^2. Escolhendo o pico de maior intensidade do espectro como pico base, cada pico complementar que pertença ao alvo é aumentado somando a sua intensidade original à intensidade do pico base vezes o fator 3, 2 ou 1 se forem classificados como pertencentes ao intervalo I, II ou III, respectivamente.

Os autores testaram retirar os complementos de peptídeos espúrios ao alvo do MS^2 ou não retirá-los. O melhor resultado foi manter todos os picos originais enquanto só os complementos do alvo foram aumentados. Mais tarde essa metodologia foi usada

no desenvolvimento do programa SuperQuant, GORSHKOV *et* al [50], acoplado ao programa comercial, Proteome Discoverer[7].

Recentemente, GORSHKOV *et* al [51], usaram os espectros multiplex modificados pelo programa SuperQuant no sequenciamento *de novo* de peptídeos, mostrando que houve um aumento de identificações entre 20 e 35%, possivelmente correspondentes aos peptídeos que tinham co-eluídos com o precursor e que não apareciam no MS^1.

Quanto maior a quantidade de fragmentos de outros peptídeos na mesma janela, mesmo em janelas de 2 *m/z* como os estudados nos trabalhos acima, mais complexo será o espectro, dificultando o sequenciamento *de novo*. Vale lembrar que é comum a ausência dos picos *b1* e *y1* nos espectros [46]. Assim, a metodologia apresentada é sempre dependente da qualidade dos espectros. A ausência no espectro de picos de fragmentos da sequência esperada (do peptídeo presente que queremos sequenciar) pode acontecer em qualquer região do espectro. Refletindo-se sobre a abordagem que cria um espectro MS^2 somente com os complementos de um precursor, a ausência de um fragmento **b** ou **y** implicaria na remoção de uma informação importante para o sequenciamento *de novo*. Contudo, o impacto destes picos faltantes na identificação de peptídeos em bancos de dados é bem reduzido.

Não há na literatura estudo sobre sequenciamento *de novo* em espectros multiplex adquiridos em janelas amplas, de 10 ou 20 *m/z*, onde, propositadamente, os peptídeos ali contidos são co-fragmentados e registrados num único espectro cuja complexidade é muito alta, diferentemente dos espectros estudados nos trabalhos anteriores em janelas de 2 *m/z*, que abordam os peptídeos co-isolados, intrusos na janela de isolamento do precursor para fragmentação. A complexidade dos espectros adquiridos em janelas amplas é muito maior que os adquiridos em janelas menores, de 2,5 *m/z*, por exemplo, gerando picos de íons fragmentos muito próximos uns dos outros, com diferenças de *m/z*, na sua maioria, menores que a tolerância de 0,02 usualmente adotada nos sequenciamentos *de novo*. Este fato dificulta ou impede a identificação dos ions-fragmentos complementares, como será analisado mais adiante.

[7] Proteome Discoverer Software da Thermo Fisher Scientific

5 DNbuilder

Na falta de um programa pré-existente disponível para iniciar o estudo aqui apresentado, um programa livre, de código aberto foi desenvolvido para resolver o problema de sequenciamento *de novo* de peptídeos, nomeado de DNbuilder, um algoritmo de formas simples. Esse programa foi implementado na plataforma Java 1.8, uma linguagem de programação amigável, robusta, moderna, de comandos simples e compatível com todos os sistemas operacionais. O computador usado para desenvolvimento do código foi um computador com processador Intel Celeron 430-1.80 GHz e 6 GB de memória RAM DDR2.

O DNbuilder processa arquivo texto com extensão MS^2 por ser uma das duas extensões mais usadas no sequenciamento *de novo*, juntamente com a extensão MGF.

Fornecendo uma visão mais ampla e geral do programa DNbuilder, inicialmente um grafo direcionado é montado a partir do espectro de massa MS^2, onde, a princípio, todos os picos do espectro tornam-se nós no grafo. Dois nós são adicionados ao grafo: um nó inicial, ou origem, representando a massa zero, e um nó final, ou destino, representando a massa do peptídeo precursor. Arestas conectam dois nós quando a diferença de massa entre dois picos resultar na massa de um resíduo de aminoácido teórico, respeitada uma tolerância δ_{max} pré-definida. O sentido da aresta é sempre do nó de menor massa para o de maior massa, percorrendo o grafo do nó inicial ao final. Neste ponto, o grafo é limpo de todos os nós e arestas sem saída, que não permitam a continuação da sequência até o nó final. Em seguida, os nós, que aqui representam picos do espectro, recebem um peso em função das intensidades de cada pico, como será explicada mais adiante. Supõem-se que os picos que constituem os nós do grafo sejam oriundos de fragmentos da série **y**. A opção de sequenciar usando a série **y**, que sequencia o peptídeo de trás para frente, foi assim decidida porque os fragmentos tipo **y** são os íons mais abundantes do espectro MS^2 por HCD [46], e consequentemente, há menos picos faltantes nesta série. Tendo o grafo e sua função de custo centrada nos nós, usou-se o algoritmo de busca DFS, examinando-se todos os caminhos que vão do nó inicial ao final.

O algoritmo de busca DSF implementado neste trabalho utilizou-se da técnica de Programação Dinâmica, que requer uma estrutura bem definida do problema e uma estrutura para armazenamento das pontuações dos subproblemas. A estrutura definida

do problema são os possíveis caminhos que saem de cada nó. A estrutura para armazenamento das pontuações dos subproblemas é um vetor que armazena o caminho que está sendo percorrido, empilhando os possíveis nós até que se encontre o nó destino. Quando isso acontece, registra-se o caminho percorrido, retorna-se ao vetor, retira-se o último nó percorrido e procura-se pelo próximo nó possível do subproblema (*backtracking*). Sempre que acontece um retorno, executa-se uma função recursiva informando o nó corrente, o nó destino e o vetor que armazena os caminhos percorridos.

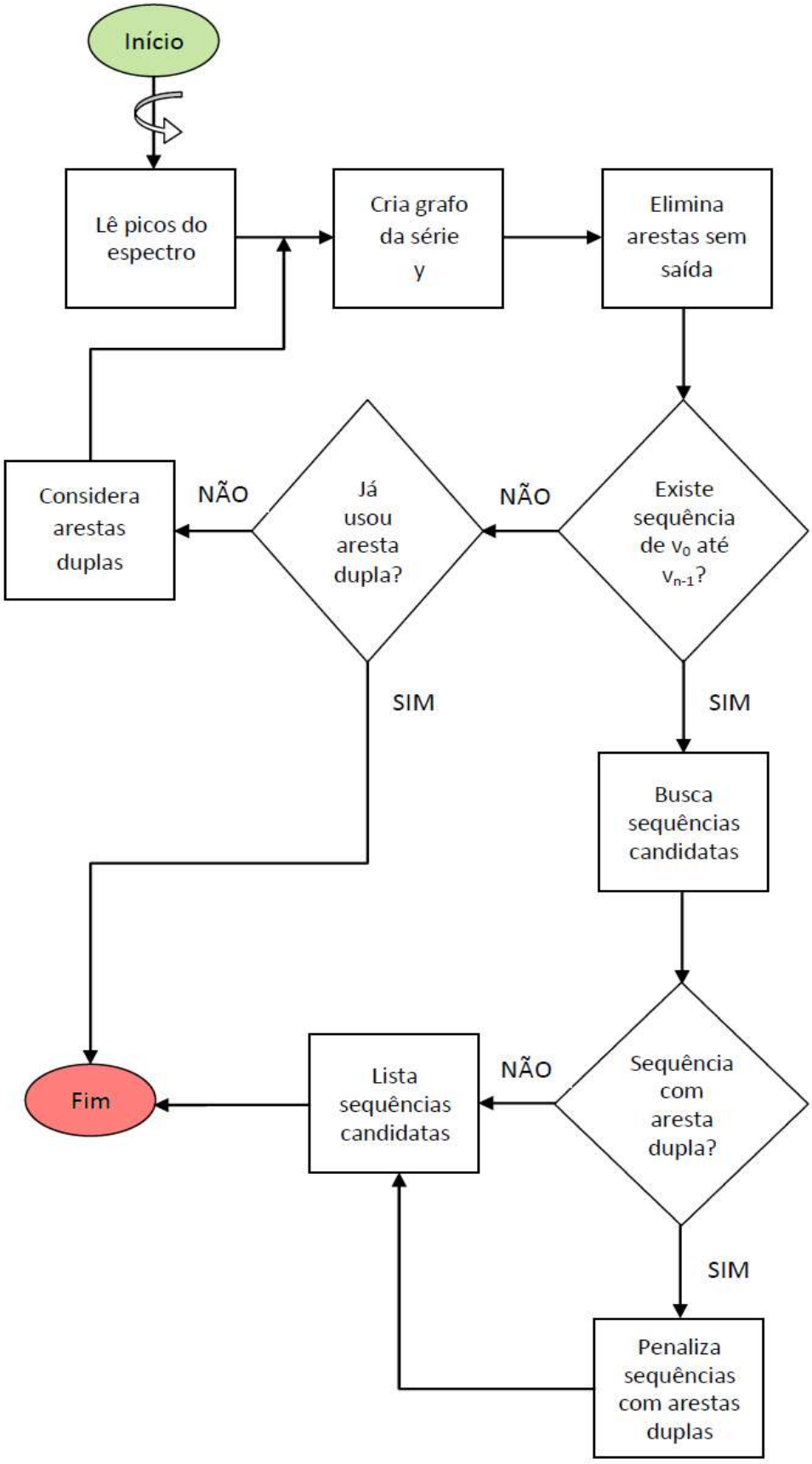


Figura 14: Fluxograma do programa DNbuilder, desde a montagem do grafo até a apresentação da lista contendo as K sequências candidatas

A pontuação final de cada caminho percorrido, cada sequência candidata, é dada pelo somatório dos pesos dos nós. Trata-se, portanto, de um problema de maximização. Ao final, apresenta-se uma lista de K sequências candidatas ordenada decrescentemente, de acordo com a pontuação geral dos caminhos. O melhor caminho maximiza a heurística da pontuação.

A Figura 14 mostra o fluxograma do programa DNbuilder, desde a montagem do grafo até a apresentação da lista contendo as K sequências candidatas. Os detalhes de cada etapa do DNBuilder, mostrado no fluxograma, será explicado a seguir.

5.1 Montagem do grafo

O sequenciamento *de novo* desenvolvido utiliza grafos, pois é a técnica que melhor descreve as características da modelagem de sequenciamento *de novo*, aliada à Programação Dinâmica para a busca dos melhores caminhos [46]. Para melhor entendimento deste trabalho, introduzimos algumas noções básicas da Teoria dos Grafos usadas neste trabalho.

Um Grafo é representado por um par (V, A), onde $V = \{v_1, \dots, v_n\}$ é um conjunto finito de n nós e $A = \{a_1, \dots, a_m\}$ é um conjunto finito de m arestas ou arcos. Os nós e arestas serão representados como $v_i \in V,\ i = 1, \dots, n,$ e $a_k \in A,\ k = (1, \dots, m)$. Uma aresta a_k é definida por um par $v_i v_j$, representando a ligação entre dois nós v_i e v_j. Se (v_i, v_j) for um par ordenado, a aresta é direcionada do nó v_i para o nó v_j. Se todos os arcos forem pares ordenados, então o grafo pode ser chamado de grafo direcionado[36].

Cada pico do espectro MS^2 será um nó do grafo, se houver uma aresta ligando ele a outro nó. Uma aresta a_k conectará dois nós (v_i, v_j) se, e somente se, a diferença entre as massas dos íons fragmentos associados a esses nós resultarem na massa de um resíduo de aminoácido conhecido[8], respeitando-se um valor de tolerância de δ_{max}, pré-determinado, condição expressa por

$$\left| m_{teórica} - \left(m(v_j) - m(v_i)\right) \right| = \delta \leq \delta_{\max} \tag{19}$$

[8] Em algumas situações, foram adotadas arestas que valessem um dipeptídeo, com a mesma tolerância δ_{max}

onde $m_{teórica}$ é a massa teórica de um dos 20 resíduos de aminoácidos naturais e $m(v_i)$ é a massa do íon fragmento associado ao nó v_i, i.é, a massa do pico do espectro MS^2 associado à v_i. O programa DNbuilder realiza os cálculos internos considerando todas as casas decimais do tipo *double* do Java, mas para efeitos de visualização, restringe em no máximo 4 casas decimais, podendo ser ajustado para uma maior quantidade.

Dois nós, v_0 e v_{n+1}, são arbitrariamente acrescentados ao grafo. O primeiro é o nó origem v_0 para representar a massa Δ para o tipo *y*, o *offset*(**y**), já definido anteriormente na Tabela 3, de massa 19,0183. O segundo é o nó destino v_{n+1} para representar a massa total do peptídeo precursor. Os caminhos no grafo definem as sequências candidatas para o espectro em análise. Os nós iniciais e finais definem o sentido do grafo direcionado da menor para a maior massa e também a série de íons-fragmentos a sequenciar, no caso a série **y**. Portanto, todos os picos do espectro são considerados, a princípio, como se fossem fragmentos do tipo **y** na montagem dos nós do grafo, já que não é possível identificá-los.

Caso não haja pelo menos um caminho do nó origem v_0 até o nó destino v_{n+1}, um novo grafo é criado admitindo tanto aresta simples, quando a diferença entre as massas de dois nós consecutivos gera um resíduo, quanto aresta dupla, quando a diferença entre as massas de dois nós resulta em qualquer combinação das massas de dois resíduos de aminoácidos. Assim, uma aresta a_k conectará dois nós (v_i, v_j) se, e somente se, a diferença entre as massas dos íons fragmentos associados a esses nós resultarem na massa de um ou dois resíduos de aminoácidos teóricos conhecidos.

Não raros são os casos de espectros que não contém o pico referente ao primeiro aminoácido da sequência esperada para a série **b (**b_1), mas contém o primeiro pico para a série **y (**y_{Nr-1}**)**. O mesmo ocorre no caso contrário, quando não existe no espectro o pico referente ao primeiro fragmento da série **y (**y_1**),** referente ao último resíduo da sequência, mas existe o ultimo para a série **b (**b_{Nr-1}**)**. Quando isso acontece pode-se inferir os picos do início e/ou do final da sequência. Desta forma, como foi usado a série **y** para sequenciar um peptídeo, se não estiverem registrados os picos dos fragmentos referentes aos resíduos de aminoácido R ou K, típicos dos peptídeos tripsinados, procura-se pelo fragmento $b_{Nr-1} = m(P) - m(K)$ ou pelo fragmento $b_{Nr-1} = m(P) - m(R)$, sendo *m(P)* a massa do precursor ou peptídeo alvo de carga +2,
para assim identificar qual resíduo finaliza a sequência.

O mesmo procedimento pode ser estendido a qualquer aresta dupla, pois é importante tentar reduzir o espaço de busca do sequenciamento *de novo* quando uma aresta dupla é adotada por causa de uma possível ausência de um pico y do caminho percorrido. Calculando-se os possíveis íons fragmentos que pudessem estar faltando através da soma da massa do pico do nó inicial da aresta dupla com cada resíduo desta aresta, é possível procurar os picos complementares para cada íon fragmento assim calculado. Caso seja encontrado apenas um dos picos complementares, a outra possibilidade deixa de ser considerada na solução dos caminhos encontrados definindo a ordem dos aminoácidos na aresta dupla. Esse procedimento evita possíveis adivinhações de caminhos na lista das sequências candidatas fornecida pelo algoritmo de busca. Quando nenhum dos dois complementos da aresta dupla é encontrado, ou encontram-se os dois complementos, todas as possibilidades de sequências encontradas pelo algoritmo de busca são apresentadas na lista.

Uma outra característica importante na montagem do grafo é que como somente são considerados os caminhos da série **y,** que vão do nó origem até o nó destino, ou seja, v_0 até v_{n+1}, não há como considerar *tags* de parte da sequência. Todas as arestas que não levam, direta ou indiretamente, ao nó v_{n+1} são eliminadas do grafo.

5.2 Função de pontuação

A chave para o sucesso da busca no grafo está em sua **função de pontuação**, ou função objetivo a ser maximizada. Essa função avalia cada caminho entre o nó origem e destino do grafo, gerando uma pontuação final para cada sequência candidata.

Observando os picos dos espectros de peptídeos trípticos usados nos testes durante o desenvolvimento do programa DNbuilder e suas intensidades, confirmou-se que as intensidades dos picos da série **y** são mais altas do que os picos da série **b**, como afirmaram MA & JOHNSON [40] e CANTÚ *et* al [17]. Esse fato é essencial para a justificativa da escolha de uma pontuação simples, adotada. Quando se maximiza as pontuações dos nós usando picos da série **b**, a importância dos complementos e evidências da série **y** são fundamentais na composição da pontuação final. As pontuações baseadas na série **b** necessitam das informações das intensidades dos picos da série **y**. Assim, adotou-se uma função de pontuação baseada apenas nas intensidades dos picos **y** cujas intensidades são naturalmente mais altas. Além disso, testou-se usar as evidências de **y** na pontuação dos nós, como será explicado no Capitulo 6, mas não foi

observada melhora nos resultados executados que justificassem sofisticar demais a pontuação. A opção mais simples foi adotada.

Para se calcular a pontuação de um nó v_i considera-se somente a intensidade de um pico do espectro associado ao nó. A Eq. (20) representa a pontuação, peso, do nó v_i.

$$C(v_i) = I(v_i) \tag{20}$$

Onde: $I(v_i)$ é a intensidade do fragmento associada ao nó;

A pontuação de um caminho PC é o somatório das pontuações dos nós do caminho, como pode ser observada na Eq. (21). O melhor caminho é definido pela maximização da função objetivo.

$$PC = \sum_{i=0}^{Nr+1} I\left(v_{f(i)}\right) \tag{21}$$

PC representa a pontuação de uma sequência candidata; $v_{f(i)}$ representa o $i^{ésimo}$ nó do caminho, que se posiciona no $f(i)^{ésimo}$ nó do grafo[9]; $I\left(v_{f(i)}\right)$ é a intensidade de um íon fragmento associado ao $i^{ésimo}$ nó do caminho do caminho que representa uma sequência de *Nr* aminoácidos.

O algoritmo de busca visa maximizar a função objetivo PC: teoricamente, quanto maior o valor da pontuação de um caminho, maior a chance desta sequência candidata ser de fato o peptídeo precursor analisado pelo espectro MS^2. Dessa forma, DNbuilder cria uma lista ordenada decrescentemente com os K maiores valores da função objetivo dos caminhos encontrados.

Um fato importante e que tem impacto na busca é o uso de arestas duplas. Quando todos os picos estão presentes no espectro, a chance de encontrar-se o caminho no grafo aumenta. É comum encontrar espectros onde nem todos os picos estão presentes, dificultando o sequenciamento *de novo*. Assim, quando não se conseguir montar um grafo para um espectro usando arestas com um resíduo de aminoácido, usou-

[9] O nó final da sequência será sempre o nó $(n+1)$ do grafo, $n+1 = f(N_r+1)$, e, consequentemente, $v_{f(N_r+1)} = v_{n+1}$. O nó inicial da sequência é sempre v_0. Então os elementos do vetor f, que determina uma sequência candidata de N_r resíduos, $f(j), \forall\, j = 0, N_r + 1$, definem os nós do grafo, ou picos do MS^2, que representam os íons-fragmentos **y** da sequência candidata, sendo os nós inicial e final fixos para todas as sequências candidatas de um grafo. Então, $f(4) = 16$ nos diz que o quarto íon-fragmento da sequência candidata, y_4, corresponde ao nó 16 do grafo.

se a combinação de dois resíduos, essencial para a busca do caminho esperado. Denomina-se, então, aresta dupla quando a aresta representa um dipeptídeo. Entretanto, sua utilização faz com que uma grande quantidade de caminhos alternativos seja criada, mesmo existindo também arestas simples no grafo. Assim, no sentido de privilegiar os caminhos mais confiáveis, com menor quantidade de arestas duplas, uma penalização é aplicada à pontuação final por cada aresta dupla do caminho.

Algumas tentativas sem sucesso para penalização foram testadas, como penalização baseada na média dos picos do espectro, ou penalização baseada no percentual da intensidade do pico mais abundante do espectro. Essas penalizações comprometiam a qualidade dos resultados, pois apesar de privilegiar alguns resultados fazendo com que a pontuação final de sequências candidatas passasse a colocá-las no topo da lista, a quantidade das perdas de sequências que antes estavam no topo da lista eram maiores.

A penalização para aresta dupla usando um percentual da pontuação final de cada sequência apresentou melhor resultado num grupo restrito de espectros usados para aferir essa penalização, não perdendo sequências candidatas já localizadas no topo da lista, e fazendo com que sequências candidatas que não se encontravam no topo da lista passassem para o topo. A dificuldade encontrada nesse caso foi a de ajustar o percentual de penalização. Os percentuais que não causaram perdas encontraram-se acima de 5% e abaixo de 12% da pontuação final do caminho. Abaixo de 5% da pontuação para penalização não alterou em nada os resultados finais. Acima de 12% fez com que houvesse perda de sequências candidatas no topo e corretas perdessem posições. O valor fixo de penalização adotado foi de 10% da pontuação como penalização para cada aresta dupla do caminho, isto é, considerando o número de arestas duplas da sequência como *NAD*, a pontuação final da sequência candidata é dada pela Eq.(22).

$$PFC = 0{,}9^{NAD} \sum_{i=0}^{Nr+1} I\left(v_{f(i)}\right) \tag{22}$$

A pontuação e penalizações adotadas aqui foram fundamentadas em testes que levaram em consideração outras possibilidades de pontuações, com e sem penalização. Os resultados serão mostrados mais adiante.

5.3 Algoritmo de busca DFS

A busca dos caminhos utilizada neste trabalho é a DFS, busca em profundidade, usando recurso da Programação Dinâmica, explicado anteriormente, pois permite a apresentação de K sequências candidatas mais pontuadas. Tarjan [43] formalizou o algoritmo DFS para busca no grafo do melhor caminho. Esse algoritmo recebe um nó origem, raiz, e a partir desse nó, progride através da avaliação dos caminhos, sempre se aprofundando, até que o nó destino seja encontrado ou até que ele se depare com um nó que não possua adjacentes (nó folha). No nosso caso, o nó folha é sempre o nó final do grafo. Então a busca retrocede (*backtracking*) e começa no próximo nó. O critério de parada do algoritmo é estabelecido quando não houver mais caminhos para serem avaliados, caracterizando uma busca exaustiva. A Figura 15 mostra o pseudocódigo de DFS para a busca do melhor caminho.

```
1:  Procedimento Principal
2:      Marcar todos os nós como não-visitados
3:      Definir uma pilha Q     // Q guarda o caminho do nó corrente até o nó inicial
4:      Definir NI o nó inicial de busca                    // NI é o nó inicial da busca
5:      executar BuscaEmProfundidade(NI)                    // Procedimento recursivo
6:  Fim Procedimento-principal
7:  Procedimento BuscaEmProfundidade(v)
8:  marcar v como visitado
9:  colocar v na pilha Q
10: Definir A como conjunto de nós adjacentes de v
11: Se A não estiver vazio
12:     para w pertencente ao conjunto de nós adjacentes A(v) fazer
13:         se w estiver como não-marcado então
14:             executar BuscaEmProfundidade (w)  // avançar na profundidade
15:         senão
16:             retornar-para-nova-avaliação          // backtracking
17:         fim-se
18:     fim-para
19: senão
20:     guarda-caminho-novo-baseado-na-PilhaQ
21:     retirar v de Q
22: fim-se
23: Fim-procedimento
```

Figura 15: Pseudocódigo de DFS para a busca do melhor caminho no grafo

Cabe aqui ressaltar que outros algoritmos de busca foram avaliados. Os algoritmos Dijkstra e A* são eficientes e rápidos para a busca de melhor caminho, sendo que o fator que mais contribuiu para a escolha do algoritmo DFS foi a facilidade de se gerar uma lista com K sequências candidatas.

Com relação à complexidade, o desempenho do algoritmo de DFS pode ser verificado como $O(n+m)$, onde n é a quantidade de nós e m indica a quantidade de arestas. Apesar de não ser a técnica mais rápida, a busca DFS é capaz de gerar de uma lista dos K melhores caminhos.

Quando todos os picos estão presentes no espectro, a chance de encontrar-se o caminho no grafo aumenta. Assim, quando não se conseguir montar um grafo para um

espectro usando arestas com um resíduo de aminoácido, aceitou-se arestas representando dipeptídeos, sendo essencial para a busca do caminho esperado. Entretanto, sua utilização faz com que uma grande quantidade de caminhos alternativos seja criada, dificultando a busca. Além disso, faltando pico da sequência no espectro, não seria possível inferir a ordem dos aminoácidos da aresta dupla na sequência, mas seria a única forma de sequenciar. Por isso é tão importante a qualidade do espectro para se obter resultados bons para o sequenciamento *de novo*.

5.4 Metodologia de sequenciamento *de novo* de peptídeos usando espectro multiplex

Os peptídeos ionizados selecionados de acordo com sua *m/z* e que se encontram dentro da janela ampla vão ter seus fragmentos registrados no mesmo espectro multiplex MS^2. Isso aumenta a complexidade na interpretação do espectro, permitindo inferir-se que quanto maior a quantidade de peptídeos em um espectro, maior a dificuldade para sequenciar o peptídeo precursor e de todos os peptídeos ali contidos. Em janelas reduzidas de 2 ou 2,5 Th, a identificação dos peptídeos presentes no espectro pode ser feita pela observação dos possíveis complementos dos fragmentos [2], se o espectro for de boa qualidade [51]. Em janelas amplas não é sempre possível a identificação dos peptídeos presentes através da procura por pares complementares. Além da complexidade na identificação dos peptídeos presentes no espectro, já que nem sempre o espectro fornece estes dados [2], outro fato importante são as grandezas das intensidades dos peptídeos ionizados presentes na janela. A princípio, peptídeos mais intensos no espectro MS^1 teriam picos dos fragmentos também mais intensos no espectro MS^2, o que facilitaria seu sequenciamento *de novo*. Em contrapartida, os peptídeos menos intensos seriam de mais difícil sequenciamento *de novo*, já que seus íons-fragmentos seriam menos intensos no espectro multiplex.

A metodologia considera a aquisição do espectro numa janela ampla, por exemplo, de 10 ou 20 *m/z*, centrada num determinado peptídeo precursor. Dentro da janela adotada, verificam-se quantos peptídeos ionizados de carga +2 são observadas e uma primeira busca de sequenciamento *de novo* para cada um destes peptídeos é executada usando o mesmo espectro multiplex original. Supõe-se o resultado da busca da sequência do peptídeo mais intenso da janela como correta, dada a maior chance de que os fragmentos sejam de fato os mais intensos. A grande questão é encontrar a

sequência dos peptídeos menos intensos, pois qualquer dos fragmentos do peptídeo mais intenso poderia confundir o algoritmo de busca. A metodologia foi desenvolvida no sentido de abater os íons-fragmentos dos peptídeos já sequenciados mais intensos para que os fragmentos dos peptídeos com menor intensidade pudessem se sobressair no espectro, facilitando o sequenciamento *de novo*.

Assim, definiu-se duas formas de descomplexação do espectro multiplex: redução drástica dos picos dos fragmentos **b** e **y** do peptídeo mais intenso, sequenciado por *de novo* e considerado como correto, e a atenuação dos picos muito acentuados que pudessem ser associados aos outros tipos de fragmentos do peptídeo mais intenso, como fragmentos tipo **a**, ou fragmentos com perdas neutras de H_2O e NH_3. Como os fragmentos não são identificáveis no espectro, um percentual dos picos mais intensos do espectro é predefinido para atenuar os picos mais intensos.

A redução drástica da intensidade dos picos do peptídeo sequenciado tipo **y** e **b** se dá pela substituição de sua intensidade pela unidade. Não foi eficiente retirá-los, pois levou a perda de informação para os outros peptídeos presentes. A atenuação da intensidade do percentual predefinido de picos que possam pertencer aos peptídeos já sequenciados se dá de forma proporcional às intensidades dos peptídeos precursores da janela e presentes no espectro de MS^1, segundo a expressão:

$$IR_{pico} = 0{,}5\, I_{pico}\, \frac{I_{ppt\ 2}}{I_{ppt\ 1}} \tag{23}$$

onde IR_{pico} é a intensidade reduzida do pico de intensidade I_{pico}, $I_{ppt\,1}$ e $I_{ppt\,2}$ são, respectivamente, a intensidade do peptídeo mais intenso, sequenciado, e menos intenso, a sequenciar. Na Eq.(23),o fator 0,5, escolhido dentre testes usando 0,25, 0,33 e 0,5, é fundamental para evitar que as intensidades dos picos dos fragmentos dos peptídeos 1 e 2 fiquem com a mesma ordem de magnitude.

O percentual de picos a serem abatidos foi definido em 33% depois de fazer testes em 20 espectros escolhidos aleatoriamente, usando abatimentos de 25, 33 e 50% dos picos mais intensos do espectro. Abater as intensidades dos 25% dos picos mais intensos do espectro permitiu que muitos picos dos fragmentos do peptídeo mais intenso continuassem com seus valores originais, confundindo a busca. Abater as intensidades de 50% dos picos mais intensos do espectro foi demasiado, pois atingiu muitos picos do peptídeo menos intenso gerando confusão na busca.

Supondo como correta a sequência encontrada no topo da lista do algoritmo de sequenciamento *de novo* do peptídeo mais intenso da janela, a heurística proposta de modificação do espectro multiplex para determinação do segundo peptídeo reduz para a unidade da intensidade dos picos dos íons-fragmentos tipo **y** do peptídeo mais intenso, exceto y1, que teve sua intensidade multiplicada pela relação Ippt2/Ippt1. As intensidades de 33% dos picos restantes mais intensos do espectro também são abatidas usando a Eq.(23).

A heurística multiplex detalhada acima foi a que apresentou melhores resultados dentre outras testadas usando espectros multiplex simulados. Nos testes que serão mostrados adiante, essa heurística foi referenciada como H4. Todas as heurísticas multiplex testadas são descritas no item 7.2, onde o resultado dos testes são apresentados.

A dinâmica para aplicação de uma heurística multiplex neste trabalho é a mesma, independentemente da quantidade de peptídeos localizados em um espectro MS^1. A Figura 16 mostra um pseudocódigo genérico para aplicação de uma heurística multiplex para peptídeos encontrados em uma janela do MS^1.

1: ProcedimentoPrincipal

2: Encontrar os *NA* peptídeos alvos de carga +2 na janela do espectro MS^1

3: Armazenar as massas e intensidades destes peptídeos alvos

4: Ordenar os peptídeos alvos decrescentemente de acordo com suas intensidades, de p_1 até p_{NA}

5: Sequenciar o peptídeo alvo p_1 da lista ordenada através do espectro MS^2

6: Para i=2, *NA*

7: Fazer cópia do espectro MS^2 original

8: Para j=1, i-1

9: Abater a sequência de p_j da cópia do MS^2 para a unidade

10: Fim-Para

11: Abater 33% dos picos mais intensos da cópia espectro MS^2 pela Eq. (23)

12: Sequenciar o peptídeo alvo p_i na cópia do espectro MS^2 deconvoluído

13: Armazenar a sequência p_i

14: Fim-Para

15: Fim-procedimentoPrincipal

Figura 16: Pseudocódigo para aplicação de heurística multiplex para uma quantidade *NA* de peptídeos de carga +2 na janela do espectro MS^1

6 METODOLOGIA E ANÁLISE DE DADOS

6.1 Espectros usados na validação das heurísticas

Dois tipos de janelas na aquisição de espectros MS^2 foram usados para testar o sequenciamento *de novo* neste trabalho: espectros MS^2 adquiridos em janelas de isolamento do precursor de 2 *m/z* e espectros MS^2 adquiridos em janelas amplas de 20 *m/z*. Para a realização dos testes deste trabalho foram utilizados diferentes conjuntos de espectros.

O primeiro conjunto é formado por um total de 64429 espectros MS^2 de precursores de carga +2, analisados pelo LTQ-Orbitrap Velos, fonte de ionização *electrospray*, e com fragmentação HCD em janelas de isolamento do precursor de 2 *m/z*. Deste conjunto 7906 espectros MS^2 foram fornecidos pela Unidade de Proteômica do departamento de Bioquímica da UFRJ. Outros 56523 espectros de MS^2 do primeiro conjunto foram fornecidos pelo Instituto Karolinska, Suécia, pelos autores da referência [6], e também utilizados recentemente no trabalho de GUTHALS, FRANK e BANDEIRA [45]. Deste conjunto foram usados nos testes 62198 espectros que puderam ser identificados pelo Comet [48]; os outros espectros foram abandonados, pois não tiveram identificação.

O segundo conjunto é constituído de 5200 espectros de MS^2 armazenados na base OPD-Open Proteomics Database [54]. Os espectros usados foram analisados por Orbitrap, com ionização por *electrospray* e fragmentação CID para peptídeos de carga +2 fornecidos por Andrew Keller *et* al [39], gerados a partir de 18 proteínas purificadas, e também referenciados pelo programa PepNovo [27]. Deste grupo 5061 espectros foram identificados pelo Comet [48] e usados nos testes do Capítulo 7.

O terceiro conjunto de espectros é proveniente da análise de uma amostra de tireóide com câncer em uma corrida DDA com janela de tamanho de 2 *m/z*, fornecidos pela Unidade de Proteômica, do Instituto de Química da Universidade Federal do Rio de Janeiro, UFRJ, contendo 17987 espectros de MS^2 de carga +2, e os respectivos espectros MS^1 da análise. Esses espectros fornecidos são de peptídeos tripsinados da amostra da tireoide , usando LTQ- Orbitrap, *electrospray* como fonte de ionização, e seleção das massas dos peptídeos mais intensos para fragmentação entre 350 e 2000 *m/z*. A energia HCD de fragmentação normalizada em 30, com a faixa inicial de registro

de íons-fragmentos de 200 *m/z*. Os espectros que puderam ser identificados pelo Comet [48] foram 16595, usados nos testes do Capitulo 7.

O quarto conjunto de espectros, usado nos testes para validação da heurística multiplex proposta, é composto de espectros provenientes da mesma amostra de tireóide com câncer, fornecidos pela Unidade de Proteômica (IQ/UFRJ), usando para a corrida o mesmo espectrômetro e a mesmas características ajustadas pelo operador dadas acima, mas sendo os espectros MS^2 adquiridos em janelas de tamanho de 20 *m/z*. Nesta análise, para cada espectro MS^1 registrado, uma quantidade máxima de 10 peptídeos foram centrados em janelas de fragmentação de 20 *m/z*, perfazendo um total de 16414 espectros MS^2 multiplex.

Um lobo de tireóide humana contendo carcinoma medular foi usado no experimento que originou o terceiro e no quarto conjunto de espectros. O tecido foi homogeneizado utilizando homogeneizador de tecido Omni e as proteínas foram extraídas em tampão de ureia 7M, tioureia 2M, 2% de sódio desoxicolato (SDC), com 50 mM de bicarbonato de trietilamônio (TEAB) na proporção de 100 mg de amostra para 5 mL de tampão. Após esse procedimento, a amostra foi incubada por 60 minutos a 4ºC e centrifugada a 20800 xg por 30 minutos a 4ºC. O sobrenadante foi precipitado *overnight* a 8ºC com acetona gelada na proporção de 1:4. Em seguida, centrifugada 20000 xg a 4ºC por 20 minutos. O precipitado foi lavado duas vezes com 0,5 mL de acetona gelada. O corpo de fundo foi resuspendido usando ureia 7M e tiureia 2M e a concentração de proteína foi medida usando Qubit©.

Para a digestão foram utilizados 100 µg de proteína total. A amostra foi incubada com ditiotreitol (DTT) em concentração final 10 mM a 30ºC por 60 minutos, foi adicionado iodocetamida (IAA) até a concentração de 40 mM e a amostra foi incubada à temperatura ambiente em ausência de luz por 30 minutos. A mistura foi diluída 10 vezes utilizando tampão HEPES 50 mM pH 8 para reduzir a concentração de ureia/tioureia para posterior digestão. A digestão foi executada incubando a solução resultante com tripsina na relação 1:50 (enzima:proteína) à temperatura de 37ºC por 16 horas. A reação foi interrompida pela adição de ácido trifluoroacético (TFA) 10% até concentração final de 0,1%.

Para a dessalinização das amostras utilizou-se a resina Poros 20 R2 de C-18, processadas em macrocolunas previamente lavadas com ACN 100% e equilibradas com TFA 0,1%. Depois as amostras foram incubadas com a resina e centrifugadas a 2000 xg

por duas vezes. A resina foi lavada duas vezes com TFA 0,1% e os peptídeos retidos foram eluídos sequencialmente com ACN 50%/TFA 0,1% e ACN 80%/TFA 0,1%, e por fim, os peptídeos foram secos em concentrador à vácuo *speedvac*.

Os peptídeos foram dissolvidos em ácido fórmico (AF) 0,1% (solvente A) foram carregadas em uma coluna guarda com fluxo de 0,5µL/minuto e separadas na coluna analítica com fluxo constante de 300 nL/minuto e um gradiente linear de 5-40% de solvente B (95% ACN, 0,1% FA) em 180 minutos. Usou-se uma ionização ESI com voltagem de 2.0kV e 200ºC no aquecedor do capilar. O espectrômetro LTQ Velos Orbitrap foi operado em modo de aquisição dependente de dados, exclusão dinâmica de 30 ms, MS^1 na faixa de 350-2000 *m/z*, resolução de 60000 (400 *m/z*), fragmentação dos 10 íons mais intensos no modo HCD, com energia de colisão normalizada de 30 e resolução 15000 (400 *m/z*) na aquisição dos espectros MS^2.

Nenhum dos espectros MS^2 usados dos quatro conjuntos de testes possuía anotações das sequências esperadas. As anotações dos conjuntos de testes usados foram obtidas a partir da identificação dos espectros através do programa Comet [48] usando o banco de dados neXtProt [44], uma plataforma para conhecimento de proteínas humanas, parametrizando a busca com contaminantes mais comuns determinados pelo próprio programa e um banco de dados *decoy* de sequências falsas. Essas sequências falsas são criadas pelo programa antes da identificação trocando a ordem dos aminoácidos das proteínas usadas a fim de dar mais confiança nos resultados. O Comet fornece uma lista de identificações com seus respectivos *scores*, ordenando-a em ordem decrescente do *score*. Quanto maior o *score*, melhor a identificação, e quanto maior a diferença entre os *scores* dos peptídeos da primeira e da segunda posição, Δ*score*, mais confiança na identificação.

6.2 Análise dos espectros

6.2.1 Identificações anotadas nos espectros

Os conjuntos de dados foram previamente analisados pelo programa Comet e seus espectros anotados e classificados por faixas de *score*. As identificações de *scores* mais altos seriam, a princípio, mais confiáveis.

O programa Comet foi usado para identificar em banco de dados os peptídeos dos espectros de cada um dos três primeiros conjuntos de testes. A Tabela 5 mostra o

número de identificações obtidas com sucesso, agrupadas por faixas dos *scores* fornecidos pelo Comet [48] para as sequências listadas na primeira posição.

Tabela 5: Número de identificações pelo COMET sobre os espectros MS^2, pontuados por faixa

Conjunto de espectros	Tipo de fragmentação	Número de MS^2 identificados pelo Comet	*Score* Comet			
			<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5
Primeiro	HCD	62198	9928	23038	14654	14578
Segundo	CID	5061	1970	1256	1089	746
Terceiro	HCD	16595	14766	1457	326	46

Observa-se que não foram identificados **2231** espectros do primeiro conjunto, **139** espectros do segundo conjunto e **1392** espectros do terceiro conjunto, representando e 3,5%, 2,7% e 7,7% dos espectros não identificados, respectivamente.

As identificações do terceiro conjunto de espectros com *score* abaixo de 1.5 chama a atenção, chegando a quase 90% do total. A fraca identificação pode indicar uma baixa qualidade destes espectros. Também é alta a identificação na faixa mais baixa de *score* para o segundo conjunto, quase 39% do total. É muito importante observar que o terceiro conjunto corresponde a uma corrida completa enquanto os outros dois conjuntos são de espectros MS^2 selecionados aleatoriamente de diversas corridas e a maioria foi usada nos testes de outros programas para sequenciamento *de novo* da literatura [5,38,46]. A princípio, estes espectros selecionados não devem conter espectros de contaminantes ou de outros elementos estranhos à amostra que aparecem comumente nos espectros iniciais das corridas, diferentemente da corrida completa do terceiro conjunto de espectros.

6.2.2 Características dos grafos

Ainda no sentido de conhecer mais os espectros dos três conjuntos usados nos testes, efetuou-se a montagem dos grafos considerando os picos do espectro como nós e arestas do tamanho de um resíduo de aminoácido, arestas simples, para obter-se o número médio de arestas, (μ), e o desvio padrão (σ). Na montagem de cada grafo usaram-se duas tolerâncias para a variação entre a massa teórica e registrada do resíduo de aminoácido, δ_{max}, uma mais e outra menos justa, a saber, 0.02 e 0.5.

Analisando melhor os espectros, observou-se a variação dos *score* das identificações. Como o Δ*score* tem muita importância na confiança da identificação,

calculou-se a média μ(Δ*score*) e o desvio padrão σ(Δ*score*) usando as duas melhores identificações do Comet de cada espectro. Cabe aqui esclarecer que os espectros que possuíam apenas uma identificação do Comet não foram contabilizados na observação das variações dos *score*, ou seja, um total de 11188, 918 e 2034 espectros do primeiro, segundo e terceiro conjunto de espectros, respectivamente, não fizeram parte dos cálculos das médias e desvios padrões. A Tabela 6 mostra a média μ(Δ*score*) e o desvio padrão σ(Δ*score*) com duas identificações por faixa de *score*.

Tabela 6: MS^2 identificados pelo programa Comet por faixa de *score* e por diferença de *score* (Δ*score*) entre o primeiro e segundo peptídeo da lista de identificações

Conjunto de teste	Faixa de *score* Comet			
	<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5
Total identificado no primeiro conjunto (62198-11188=51010)	7207	18457	12597	12749
Média: μ(Δ*score*)	0,07	0,15	0,40	1,10
Desvio: σ(Δ*score*)	0,12	0,09	0,17	0,27
Total identificado no segundo conjunto (5061-918=4143)	519	1586	1084	954
Média: μ(Δ*score*)	0,06	0,14	0,41	1,10
Desvio: σ(Δ*score*)	0,07	0,13	0,28	0,51
Total identificado no terceiro conjunto (16595-2034=14561)	12835	1369	313	44
Média: μ(Δ*score*)	0,11	0,67	1,10	1,57
Desvio: σ(Δ*score*)	0,14	0,23	0,26	0,31

Os resultados mostram que as distâncias médias entre os *scores* variaram muito pouco em todos os conjuntos de testes, mostrando certa uniformidade nos *score* das identificações dentro de cada faixa de *score*.

Outro teste feito avalia a média de nós e arestas e seus respectivos desvios usando os espectros originais dos conjuntos de testes. A Tabela 7, 8 e 9 mostram o número médio de nós e arestas (μ) e o respectivo desvio padrão (σ) nos grafos, agrupados por faixa de *score* do Comet, para o primeiro, segundo e terceiro conjunto de espectros, respectivamente. Note que nem todos os espectros identificados geram os grafos conexos ligando o nó inicial ao final. Por este motivo, o número de espectros identificados é diferente do número de espectros que geram grafos com pelo menos um caminho completo do nó inicial ao final, como mostrado nestas tabelas.

Tabela 7: Número médio de nós e arestas (μ), e respectivo desvio padrão (σ), para os grafos gerados pelos espectros MS^2 do conjunto, por faixa de *score* (Comet) para o **primeiro conjunto de espectros**

Total de espectros MS^2	Por *score* Comet			
	<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5
Tolerância 0,02				
62198	954	4979	3760	1986
μ Nós (σ Nós)	8 (1,72)	10 (2,37)	10 (2,22)	13 (3,32)
μ Arestas (σ Arestas)	8 (2,64)	10 (3,63)	10 (3,47)	12 (5,35)
Tolerância 0,5				
62198	3122	11126	7488	4356
μ Nós (σ Nós)	28 (13,46)	34 (16,91)	39 (19,12)	54 (29,92)
μ Arestas (σ Arestas)	40 (23,22)	51 (30,16)	60 (34,26)	85 (56,42)

Tabela 8: Número médio de nós e arestas (μ), e respectivo desvio padrão (σ), para os grafos gerados pelos espectros MS^2 do conjunto, por faixa de *score* (Comet) para o **segundo conjunto de espectros**

Total de espectros MS^2	Por score Comet			
	<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5
Tolerância 0,02				
5061	74	381	294	138
μ Nós (σ Nós)	10 (1,99)	10 (2,38)	11 (2,16)	12 (1,93)
μ Arestas (σ Arestas)	10 (3,69)	11 (4,14)	11 (3,77)	12 (3,35)
Tolerância 0,5				
5061	285	955	648	320
μ Nós (σ Nós)	29 (13,76)	32 (15,57)	31 (13,53)	31 (11,75)
μ Arestas (σ Arestas)	42 (23,55)	48 (27,77)	45 (23,96)	44 (21,08)

Tabela 9: Número médio de nós e arestas (μ), e respectivo desvio padrão (σ), para os grafos gerados pelos espectros MS^2 do conjunto, por faixa de *score* (Comet) para o **terceiro conjunto de espectros**

Total de espectros MS^2	Por *score* Comet			
	<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5
Tolerância 0,02				
16595	1359	189	27	2
μ Nós (σ Nós)	11 (4,81)	13 (3,97)	14 (2,71)	17 (0,50)
μ Arestas (σ Arestas)	12 (8,72)	13 (6,37)	14 (4,11)	17 (0,50)
Tolerância 0,5				
16595	2703	365	62	10
μ Nós (σ Nós)	54 (34,10)	63 (33,23)	75 (36,07)	70 (33,43)
μ Arestas (σ Arestas)	86 (63,75)	102 (64,17)	123 (70,15)	116 (68,21)

O uso de uma tolerância muito justa, 0,02, faz com que a média dos nós e arestas dos grafos sejam baixas, o que diminui a complexidade do grafo, facilitando a busca da sequência correta. Já com uma tolerância mais alta, 0,5, essas médias se elevam acentuadamente, permitindo o aparecimento de caminhos alternativos na busca, inclusive caminhos mais bem pontuados que a sequência esperada, dificultando a busca.

Como já foi dito anteriormente, identificar no Comet não quer dizer que os caminhos do peptídeo identificado estejam presentes no grafo, isto é, que todos os fragmentos **y** dos peptídeos identificados tenham sido registrados no espectro. Considerando somente o primeiro peptídeo válido da lista fornecida pelo Comet, o de maior *score*, contou-se a quantidade de grafos deste conjunto de espectros de MS^2 que contém o caminho completo (representando a sequência completa de aa) do peptídeo identificado em primeiro lugar pelo Comet. Surpreendentemente, o grupo que concentra a maior quantidade de caminhos completos é o de *score* entre 1,5 e 2 nos dois primeiros conjuntos de espectros.

O primeiro e o segundo conjuntos de espectros mostram maior quantidade de espectros com maiores médias de nós e arestas na faixa entre [1,5 e 2), com uma distribuição entre as outras faixas não muito distantes.

O terceiro conjunto mostra-se muito diferente, com uma concentração de espectros com maiores médias de nós e arestas na faixa < 1,5. Essa concentração chama a atenção pela distância das outras faixas. Para a tolerância 0,02, na faixa <1,5 são 1359 espectros, na faixa entre [1,5 e 2) são 189 espectros, na faixa entre [2 e 2,5) são 27 espectros e na faixa >=2,5 apenas 2. Usando a tolerância 0,5 Da seguiu-se uma distribuição parecida. Nesse conjunto de espectros a faixa predominante é a de *score* menor que 1,5. Esse fato reforça a suspeita de que os espectros não são de boa qualidade. Um fato que pode justificar essa baixa qualidade dos espectros é o fato de que a massa inicial dos fragmentos de todos os espectros desse terceiro conjunto inicia em 200 m/z, sendo que o primeiro fragmento do tipo **y**, lisina ou arginina, mesmo somando o *offset* do tipo **y**, tem a massa menor que 200Da, e não aparecem no espectro. A massa do pico da série **y** para a lisina é 147,1132Da (128,0949+19,0183), enquanto que a massa do pico da série y para a arginina é 175,1194Da (156,1011+19,0183). Isso dificulta muito o sequenciamento por *de novo*. Para possibilitar usar arestas simples neste conjunto de teste foi necessário inserir artificiamente nós relativos aos picos referentes à lisina e arginina.

As características analisadas constatam que os grafos são reduzidos para tolerância de 0,02 Da, a mais adotada pelos algoritmos para sequenciamento *de novo*, especialmente quando a análise dos peptídeos é feita em alta resolução. Para tolerâncias maiores, os grafos ficam mais densos com muitas arestas oriundas de fragmentos que não fazem parte do trajeto procurado podendo ainda possuir intensidades elevadas. Porque as respostas para tolerâncias menores são melhores, o que será visto adiante, não faria sentido tentar reduzir, através de podas, nós ou picos dos espectros. De fato, toda tentativa executada para reduzir o número de picos do espectro na esperança de retirar ruídos ou fragmentos espúrios, não foi eficiente, provocando perdas de informação importantes e consequente redução dos sequenciamentos *de novo* corretos, mesmo para peptídeos analisados em baixa resolução, onde a maioria dos algoritmos sugere a tolerância de 0,5 Da.

6.2.3 Montagem dos grafos versus sequências identificadas

Em muitos espectros foram observadas ausências de picos, principalmente os picos das extremidades, de maiores e menores massas. Não raros são os casos de espectros que não contém o pico referente ao primeiro resíduo de aminoácido da sequência esperada para a série **b**, mas contém o primeiro pico para a série **y**. Também

pode ocorrer o caso contrário, quando não existe no espectro o pico referente ao primeiro aminoácido para a série **y** (último resíduo da sequência), mas existe para a série **b**.

Vale lembrar que os grafos montados consideram todos os picos dos espectros. Os nós soltos no grafo, com grau de entrada ou saída zero e que não sejam o nó inicial ou final do grafo, são retirados, assim como as arestas a eles ligadas.

Para avaliar a presença de pelo menos um trajeto completo no grafo, investigou-se também a inserção de nós (ou picos no espectro) nos extremos da sequência e também a criação de arestas duplas. A inserção de novos picos nas extremidades do espectro aumenta consideravelmente o grafo já que teríamos que usar as 20 massas dos aminoácidos teóricos, chegando a uma média próxima de 300 picos inseridos no espectro somente combinando dois aminoácidos. Apesar de não garantir que a sequência esperada esteja presente, já que podem estar faltando também outros fragmentos, essa inserção aumentou a presença das sequências esperadas em mais de 50% dos espectros, como pode ser verificado nas Tabelas 10 e 11.

Importante lembrar que pelo fato da sequência estar presente no espectro modificado não significa que o sequenciamento por *de novo* encontrará o caminho. Mas também é fato que, para que o sequenciamento *de novo* tenha chance de sucesso, é necessário que os fragmentos sejam completos ou quase no espectro para gerar também grafos com caminhos completos.

As Tabelas 10 e 11 mostram quantos espectros do primeiro e segundo conjunto, respectivamente, com sequências identificadas pelo Comet, contém os caminhos esperados no grafo, trajeto da primeira identificação, para diferentes montagens de grafos, agrupando-os por faixa de *score* Comet. As tabelas também mostram a média do número de nós e a de arestas dos grafos montados a partir dos espectros MS^2. Nas duas tabelas utilizaram diferentes montagens de grafo: grafos usando os dados originais dos espectros, três diferentes grafos com a inclusão de um, dois e três potenciais picos ausentes nas extremidades dos espectros (ligados aos fragmentos y_0 e y_n, isto é., ligados aos nós inicial e final do grafo) e grafos usando arestas duplas em três diferentes situações. A primeira, arestas duplas ligadas somente aos fragmentos y_0 e y_n do espectro original, ou seja, nas extremidades do espectro; a segunda, arestas duplas usadas em todo o espectro original, não apenas nas extremidades; a terceira e última situação, inclusão de 1 pico nas extremidades e arestas duplas em todo o espectro. Foram usadas

três diferentes tolerâncias no cálculo do resíduo para criação das arestas do grafo: 0,02, 0,05 e 0,5 Da.

Tabela 10: Número de espectros do primeiro conjunto com identificação pelo Comet que contém no respectivo grafo todo o caminho esperado (MS^2 que contém todos os picos dos fragmentos teóricos esperados), considerando a inclusão até três picos e a combinação de aminoácidos nos flancos e em todo o espectro

Teste (tolerância = 0,02 Da)	Média de nós	Média de arestas	Faixa de *score* Comet				Total
			<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5	
Espectros originais	11	13	752	4399	4467	3352	12970
Espectros com inclusão de 1 pico nas duas extremidades	15	19	1524	7118	7637	7243	23522
Espectros com inclusão de 2 picos nas duas extremidades	39	102	1929	8002	8327	8834	27092
Espectros com inclusão de 3 picos nas duas extremidades	203	1302	2609	9545	8907	9722	30783
Combinação de até dois aminoácidos nas arestas nas duas extremidades do espectro	24	79	1524	7118	7637	7243	23522
Combinação de até dois aminoácidos nas arestas em todo o grafo	24	143	1931	8214	8467	8901	27513
Inclusão de um pico nas duas extremidades e combinação de até dois aminoácidos nas arestas em todo o grafo	41	188	2252	9293	8685	8912	29142

Teste (tolerância = 0,05 Da)	Média de nós	Média de arestas	Faixa de *score* Comet				Total
			<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5	
Espectros originais	18	28	789	4586	4628	3584	13587
Espectros com inclusão de 1 pico nas duas extremidades	23	37	1621	7450	7853	7700	24624
Espectros com inclusão de 2 picos nas duas extremidades	57	201	2115	8480	8576	9332	28503
Espectros com inclusão de 3 picos nas duas extremidades	209	1738	3037	10503	9292	10150	32982
Combinação de até dois aminoácidos nas arestas nas duas extremidades do espectro	39	229	1621	7450	7853	7700	24624
Combinação de até dois aminoácidos nas arestas em todo o grafo	41	483	2221	8497	8589	9363	28670
Inclusão de um pico nas duas extremidades e combinação de até dois aminoácidos nas arestas em todo o grafo	45	559	2971	9812	9896	9488	32167

Teste (tolerância = 0,5 Da)	Média de nós	Média de arestas	Faixa de *score* Comet				Total
			<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5	
Espectros originais	32	68	838	4788	4800	3667	14093
Espectros com inclusão de 1 pico nas duas extremidades	37	79	1698	7759	8041	7820	25318
Espectros com inclusão de 2 picos nas duas extremidades	77	312	2290	8985	8854	9498	29627
Espectros com inclusão de 3 picos nas duas extremidades	215	1746	3551	11460	9733	10322	35066
Combinação de até dois aminoácidos nas arestas nas duas extremidades do espectro	57	361	1698	7759	8041	7820	25318
Combinação de até dois aminoácidos nas arestas em todo o grafo	61	631	2315	8522	8617	9499	28953
Inclusão de um pico nas duas extremidades e combinação de até dois aminoácidos nas arestas em todo o grafo	65	894	2990	9931	10931	9514	33366

Observa-se nos resultados mostrados na Tabela 10 usando a tolerância 0,02 o fato de 6,4% do total de espectros dos espectros possuírem picos faltantes no meio do espectro (y_{n-3}, y_{n-3}etc até y_3 – correspondentes a b_3, b_4, ..., b_{n-3}), ou seja, 3991 = (27513-

23522) dividido por 62198. Já para a tolerância 0,5, o percentual tem pequena variação para 5,9% do total de espectros. A falta dos fragmentos iniciais e finais do espectro também é alta. A adoção de aresta dupla aumenta muito a presença de trajetos das sequencias anotadas nos grafos, e sem a adoção de picos artificiais indesejáveis.

Outro fato observado é o de que o considerável ganho na presença do caminho com a inclusão de picos no espectro é muito inferior proporcionalmente ao aumento da complexidade, determinada pela quantidade de nós e arestas do grafo.

A adoção de aresta dupla, permitindo a representação de dipepídeos em todo o espectro, parece ser o passo mais seguro para o sequenciamento *de novo*, pois não se introduz nós falsos no espectro, já que nós falsos, mesmo com pontuação nula, aumentaria a incerteza das sequências após a busca.

Tabela 11: Número de espectros do segundo conjunto com identificação pelo Comet que contém no respectivo grafo (completo) todo o caminho esperado (MS^2 que contém todos os picos dos fragmentos teóricos esperados), considerando a inclusão até três picos e a combinação de aminoácidos nos flancos e em todo o espectro

Teste (tolerância = 0,02 Da)	Média de nós	Média de arestas	Faixa de *score* Comet				Total
			<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5	
Espectros originais	11	12	56	308	322	250	936
Espectros com inclusão de 1 pico nas duas extremidades	14	18	116	524	607	530	1777
Espectros com inclusão de 2 picos nas duas extremidades	39	103	138	606	665	654	2063
Espectros com inclusão de 3 picos nas duas extremidades	147	1031	188	762	722	727	2399
Combinação de até dois aminoácidos nas arestas nas duas extremidades do espectro	13	59	116	524	607	530	1777
Combinação de até dois aminoácidos nas arestas em todo o grafo	39	341	133	535	684	621	1973
Inclusão de um pico nas duas extremidades e combinação de até dois aminoácidos nas arestas em todo o grafo	44	441	152	554	699	627	2032

Teste (tolerância = 0,05 Da)	Média de nós	Média de arestas	Faixa de *score* Comet				Total
			<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5	
Espectros originais	17	26	58	328	345	262	993
Espectros com inclusão de 1 pico	21	34	124	563	629	572	1888
Espectros com inclusão de 2 picos	56	198	155	654	684	691	2184
Espectros com inclusão de 3 picos	171	1473	237	844	760	756	2597
Combinação de até dois aminoácidos nas arestas nas duas extremidades do espectro	20	91	124	563	629	572	1888
Combinação de até dois aminoácidos nas arestas em todo o grafo	44	497	143	592	649	688	2072
Inclusão de um pico em cada extremidade e combinação de até dois aminoácidos nas arestas em todo o grafo	58	591	165	635	677	692	2169

Teste (tolerância = 0,5 Da)	Média de nós	Média de arestas	Faixa de *score* Comet				Total
			<1,5	[1,5 e 2)	[2 e 2,5)	>=2,5	
Espectros originais	29	61	63	345	365	267	1040
Espectros com inclusão de 1 pico	34	69	131	589	651	580	1951
Espectros com inclusão de 2 picos	74	300	171	699	719	701	2290
Espectros com inclusão de 3 picos	174	1593	276	925	796	765	2762
Combinação de até dois aminoácidos nas arestas nas duas extremidades do espectro	32	128	131	589	651	580	1951
Combinação de até dois aminoácidos nas arestas em todo o grafo	53	763	155	632	722	602	2111
Inclusão de um pico em cada extremidade e combinação de até dois aminoácidos nas arestas em todo o grafo	62	883	177	662	757	729	2325

A Tabela 11 mostra resultados semelhantes aos da Tabela 10 reforçando as características dos espectros de fragmentos CID e HCD. Para tolerância 0,02 Da, 4,8% do total de espectros possuem picos faltantes no meio do espectro, e para a tolerância 0,5 Da, o percentual se cai para 3,6% do total de espectros, resultado menor que os observados no primeiro conjunto. Interessante observar que em ambos os conjuntos, adotar arestas duplas foi muito mais eficiente que sair introduzindo nós dos possíveis

fragmentos faltantes, constituindo uma representação mais próxima da realidade, já que se atém às informações contidas nos espectros.

A prática de combinação de resíduos de aminoácidos para completar possíveis *gap* foi adotada por alguns programas para sequenciamento *de novo* [25-28,30-33].

Observa-se, mais uma vez, a grande dependência do sequenciamento *de novo* na qualidade dos espectros, pois depende do registro de picos de todos os fragmentos necessários para tornar o sequenciamento *de novo* possível. Pode-se afirmar, então, que considerando uma tolerância de 0,02 Da na montagem dos grafos, dos 62198 espectros HCD identificados no Comet, 12970 espectros, correspondente a 20,8% das identificações do conjunto, são teoricamente passíveis de serem sequenciados usando grafos com aresta simples, e 27513 espectros, correspondente a 44,2% das identificações, se admitidas arestas duplas nos grafos. Quanto aos espectros CID, dos 5061 identificados, 936 espectros, correspondentes a 18,5% das identificações, poderiam ser sequenciados usando grafos de aresta simples, e 1973 espectros, correspondentes a 38,9% das identificações, usando grafos de aresta dupla. Como pode ser observado nas Tabelas 10 e 11, estas seriam as expectativas máximas de sucesso possíveis nos testes com qualquer sequenciamento *de novo* usando estes conjuntos de dados e cuja heurística para o sequenciamento *de novo* opte pelos fragmentos tipo **y**. Seria, portanto, a expectativa máxima para o DNbuilder.

6.3 Montagem da pontuação no DNbuilder

A busca pelo melhor caminho do grafo depende da pontuação, que é o ponto chave para o sucesso do sequenciamento *de novo* de peptídeo. O que se almeja é que o melhor caminho encontrado pela busca seja a sequência do peptídeos, ou seja, a reconstrução dos resíduos de aminoácidos do peptídeo precursor. Assim, uma boa pontuação de cada nó do grafo é de extrema importância para a solução do problema de sequenciamento *de novo*.

O que se pretende neste teste é avaliar a influência das evidências do fragmento tipo **y**, isto é, **b, b**-H_2O**, b**-NH_3**, a, a**-H_2O**, a**-NH_3**, y, y**-H_2O **e y**-NH_3, no resultado da metodologia para sequenciamento *de novo* do DNbuilder proposta, usando diferentes montagens de grafos e tolerância 0,02 Da, a saber:

1. Usando somente os picos originais dos espectros e arestas de um resíduo de aminoácido (chamadas neste trabalho de arestas simples);
2. Criação de arestas de até 2 resíduos (chamadas neste trabalho de arestas duplas) para a montagem de todo o grafo;
3. Arestas duplas em todo o grafo usadas somente quando não houver caminho do nó origem até o nó destino para grafos de aresta simples;

Os espectros usados nos testes foram o primeiro, segundo e terceiro conjuntos de espectros descritos anteriormente. Em cada uma das três diferentes montagens de grafo elencadas acima, considerou-se diferentes pontuações, chamadas de PT_x, dos nós do grafo para avaliar o impacto das evidências dos picos na pontuação final da busca, considerando todos os picos como se fossem fragmentos do tipo **y**, pois a série **y** é a usada no sequenciamento *de novo* do algoritmo DNbuilder. São elas:

PT1. Intensidades dos picos **y** como único peso dos nós;
PT2. Somatório das intensidades dos picos **y** e **b** como peso dos nós;
PT3. Somatório das intensidades dos picos **y**, **b** e **a** como peso dos nós;
PT4. Somatório das intensidades dos picos **y**, **b**, **a** e perdas de H_2O ;
PT5. Somatório das intensidades dos picos **y**, **b**, **a**, e perdas de NH_3;
PT6. Somatório das intensidades dos picos **y**, **b**, **a**, e perdas de H_2O e NH_3;
PT7. Somatório das intensidades dos picos **y**, **b**, **a**, **b2** e **y2** e perdas de H_2O e NH_3;
PT8. Somatório das intensidades dos picos **y**, **b**, **a**, **b2**, **y2** e perdas H_2O e NH_3, e perdas duplas de H_2O-H_2O e H_2O-NH_3 como peso dos nós.

Ainda foi considerada a adição da penalização na pontuação final das sequências candidatas quando admitida a aresta dupla, que seria modificadora da posição da sequência na lista de sequências candidatas fornecidas pelo algoritmo de busca, na tentativa de melhorar a quantidade de sequenciamentos *de novo* corretos na primeira posição. Nos testes aqui apresentados, o $I(v_{f(i)})$ da Eq. (22) que definiu a pontuação final com penalização do DNbuilder, representará a pontuação de cada nó $v_{f(i)}$ do caminho para cada pontuação listada acima.

Nos testes abaixo, executou-se o programa DNbuilder usando as três diferentes montagens de grafo combinadas com as oito formas de pontuação apresentadas acima. Quando o grafo admitia aresta dupla, também foram gerados resultados considerando a penalização após a busca das sequencias em cujos trajetos havia aresta dupla.

A sequência mais bem pontuada de cada lista foi comparada com a respectiva identificação do espectro feita em banco de dados usando o programa Comet. Só foram considerados corretos os sequenciamentos *de novo* que colocaram a sequência identificada pelo Comet no topo da lista. Os resultados com aresta dupla foram considerados corretos quando a aresta dupla da sequência representou o dipeptídeo presente, mesmo não sendo possível definir a ordem deles na sequência. Os resultados foram agrupados por conjunto de espectro usado, montagem do grafo utilizada e os picos considerados na pontuação.

A Tabela 12 mostra a quantidade de sequenciamentos *de novo* corretos do DNbuilder para cada uma das três montagens diferentes de grafo usando as oito formas de pontuar os nós do grafo considerando as evidências, separadas por *score* Comet, **usando o primeiro conjunto de espectros**. Essa tabela também mostra as quantidades de sequências candidatas que deixaram o topo da lista e as que foram para o topo da lista, tendo como base de comparação os resultados usando somente a intensidade dos fragmentos tipo **y** na pontuação dos nós.

Tabela 12: Comparação entre quantidades de grafos (espectros) sequenciados corretamente pelo DNbuilder na primeira posição da lista, em diferentes modelagens do grafo e diferentes evidências na pontuação dos nós, **usando espectros do primeiro conjunto**. Os melhores resultados foram marcados para as diferentes evidências por cada faixa do *score* Comet

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **arestas simples**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	290	2729	2968	2021	8008
PT2	286	2715	2960	2009	7970
PT3	284	2708	2956	2002	7950
PT4	286	2676	2928	1982	7872
PT5	283	2722	2979	1998	7982
PT6	284	2696	2947	1990	7917
PT7	283	2689	2932	1986	7890
PT8	283	2671	2922	1978	7854

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado.		
	Total	Saíram	Entraram
PT1	8008	---	---
PT2	7970	86	48
PT3	7950	109	51
PT4	7872	231	95
PT5	7982	154	128
PT6	7917	224	133
PT7	7890	252	134
PT8	7854	300	146

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **arestas duplas**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	446	3081	3609	3046	10182
PT2	447	3016	3511	2893	9867
PT3	444	3000	3482	2841	9767
PT4	435	2932	3378	2742	9487
PT5	437	3031	3498	2791	9757
PT6	432	2968	3417	2735	9552
PT7	429	2956	3395	2713	9493
PT8	420	2897	3285	2633	9235

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado.		
	Total	Saíram	Entraram
PT1	10182	---	---
PT2	9867	488	173
PT3	9767	589	174
PT4	9487	958	263
PT5	9757	751	326
PT6	9552	990	360
PT7	9493	1053	364
PT8	9235	1317	370

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **aresta dupla apenas quando aresta simples não criar caminho do nó origem até o nó destino**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	503	3626	4176	3600	11905
PT2	503	3601	4138	3481	11723
PT3	500	3586	4124	3444	11654
PT4	498	3530	4047	3358	11433
PT5	496	3610	4135	3406	11647
PT6	491	3562	4062	3360	11475
PT7	488	3551	4049	3349	11437
PT8	483	3514	4022	3299	11318

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado.		
	Total	Saíram	Entraram
PT1	11905	---	---
PT2	11723	312	130
PT3	11654	384	133
PT4	11433	686	214
PT5	11647	522	264
PT6	11475	717	287
PT7	11437	760	292
PT8	11318	900	313

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **aresta dupla, com pontuação final penalizada em 10% da pontuação da sequência para cada aresta dupla do caminho**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	538	3796	4555	4229	13118
PT2	533	3735	4503	4149	12920
PT3	532	3717	4487	4109	12845
PT4	528	3661	4413	4018	12620
PT5	530	3733	4507	4090	12860
PT6	525	3696	4443	4025	12689
PT7	523	3694	4435	4026	12678
PT8	517	3661	4404	3997	12579

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	13118	---	---
PT2	12920	361	163
PT3	12845	444	171
PT4	12620	747	249
PT5	12860	565	307
PT6	12689	753	324
PT7	12678	779	339
PT8	12579	896	357

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **aresta dupla apenas quando aresta simples não criar caminho do nó origem até o nó destino, com a penalização de 10% da pontuação da sequência para cada aresta dupla do caminho**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	536	3785	4531	4221	13073
PT2	533	3761	4516	4169	12979
PT3	532	3748	4501	4134	12915
PT4	527	3696	4440	4045	12708
PT5	528	3767	4515	4109	12919
PT6	525	3728	4454	4051	12758
PT7	523	3724	4445	4048	12740
PT8	520	3688	4425	4023	12656

Primeiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	13073	---	---
PT2	12979	216	122
PT3	12915	288	130
PT4	12708	564	199
PT5	12919	411	257
PT6	12758	581	266
PT7	12740	613	280
PT8	12656	716	299

Para o primeiro conjunto de espectros os resultados mostram que, para as cinco diferentes montagens de grafo, usar as intensidades dos picos **y** na pontuação dos nós é mais eficiente na busca dos caminhos não justificando complicar a pontuação. Tendo como base as sequências corretas usando somente os picos **y** para pontuação dos nós, em todos os outros tipos de pontuações ocorreram de sequências corretas perderem o topo da lista, enquanto outras ganharem o topo. Em todas as formas de pontuação houve mais perdas do que ganho, comprovando a eficiência de uma pontuação simples baseada na série **y**.

As Tabelas 13 e 14 mostram os resultados para o segundo e o terceiro conjunto de testes.

Tabela 13: Comparação entre quantidades de grafos (espectros) sequenciados corretamente pelo DNbuilder na primeira posição da lista, em diferentes modelagens do grafo e diferentes evidências na pontuação dos nós, **usando espectros do segundo conjunto**. Os melhores resultados foram marcados para as diferentes evidências por cada faixa do *score* Comet

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **arestas simples**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	21	186	218	144	569
PT2	20	183	221	142	566
PT3	20	183	222	142	567
PT4	20	182	218	139	559
PT5	20	183	222	141	566
PT6	20	185	219	138	562
PT7	20	182	216	138	556
PT8	19	182	214	139	554

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	569	---	---
PT2	566	6	3
PT3	567	6	4
PT4	559	18	8
PT5	566	10	7
PT6	562	15	8
PT7	556	21	8
PT8	554	23	8

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **arestas duplas**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	35	207	296	233	771
PT2	36	201	286	218	741
PT3	36	201	286	216	739
PT4	33	198	281	207	719
PT5	35	201	288	206	730
PT6	33	199	282	207	721
PT7	34	195	278	206	713
PT8	33	195	273	205	706

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	771	---	---
PT2	741	46	16
PT3	739	48	16
PT4	719	76	24
PT5	730	67	26
PT6	721	82	32
PT7	713	89	31
PT8	706	94	29

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **aresta dupla apenas quando aresta simples não criar caminho do nó origem até o nó destino**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	38	247	339	270	894
PT2	38	245	338	262	883
PT3	38	245	339	261	883
PT4	36	242	330	254	862
PT5	38	244	341	255	878
PT6	36	243	333	254	866
PT7	37	239	330	254	860
PT8	35	239	327	257	858

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	894	---	---
PT2	883	26	15
PT3	883	27	16
PT4	862	53	21
PT5	878	40	24
PT6	866	54	26
PT7	860	59	25
PT8	858	62	26

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **aresta dupla penalizada em 10% da pontuação da sequência para cada aresta dupla do caminho**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	41	252	383	300	976
PT2	41	251	379	293	964
PT3	41	252	380	291	964
PT4	38	249	369	286	942
PT5	41	248	379	289	957
PT6	38	251	372	284	945
PT7	39	247	370	283	939
PT8	38	247	370	284	939

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	976	---	---
PT2	964	26	14
PT3	964	28	16
PT4	942	55	21
PT5	957	42	23
PT6	945	56	25
PT7	939	60	23
PT8	939	62	25

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **aresta dupla apenas quando aresta simples não criar caminho do nó origem até o nó destino, com a penalização de 10% da pontuação da sequência para cada aresta dupla do caminho**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	41	255	384	298	978
PT2	41	256	385	295	977
PT3	41	257	386	293	977
PT4	39	254	375	288	956
PT5	41	256	385	291	973
PT6	39	256	378	286	959
PT7	40	253	375	285	953
PT8	39	252	376	286	953

Segundo conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	978	---	---
PT2	977	14	13
PT3	977	16	15
PT4	956	42	20
PT5	973	27	22
PT6	959	41	22
PT7	953	46	21
PT8	953	49	24

Tabela 14: Comparação entre quantidades de grafos (espectros) sequenciados corretamente pelo DNbuilder na primeira posição da lista, em diferentes modelagens do grafo e diferentes evidências na pontuação dos nós, **usando espectros do terceiro conjunto**. Os melhores resultados foram marcados para as diferentes evidências por cada faixa do *score* Comet

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **arestas simples**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	375	180	27	4	586
PT2	377	181	27	4	589
PT3	372	179	26	4	581
PT4	355	168	25	4	552
PT5	375	179	28	5	587
PT6	370	176	28	4	578
PT7	370	173	28	4	575
PT8	367	171	28	5	571

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	586	---	---
PT2	589	2	5
PT3	581	11	6
PT4	552	42	8
PT5	587	14	15
PT6	578	20	12
PT7	575	23	12
PT8	571	28	13

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **arestas duplas**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	468	196	24	7	695
PT2	460	199	21	7	687
PT3	433	190	19	7	649
PT4	401	171	18	5	595
PT5	437	198	21	7	663
PT6	421	191	20	6	638
PT7	410	183	19	6	618
PT8	413	178	20	6	617

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	695	---	---
PT2	687	23	15
PT3	649	61	15
PT4	595	118	18
PT5	663	72	40
PT6	638	91	34
PT7	618	104	27
PT8	617	105	27

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **aresta dupla apenas quando aresta simples não criar caminho do nó origem até o nó destino**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	646	306	42	7	1001
PT2	640	309	40	6	995
PT3	609	298	37	6	950
PT4	577	276	36	5	894
PT5	613	303	41	7	964
PT6	602	293	40	5	940
PT7	599	286	39	5	929
PT8	599	280	38	6	923

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	1001	---	---
PT2	995	21	15
PT3	950	64	13
PT4	894	126	19
PT5	964	74	37
PT6	940	92	31
PT7	929	99	27
PT8	923	106	28

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **aresta dupla penalizada em 10% da pontuação da sequência para cada aresta dupla do caminho**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	724	338	55	8	1125
PT2	728	349	54	8	1139
PT3	705	337	50	8	1100
PT4	667	316	48	7	1038
PT5	714	347	52	9	1122
PT6	695	341	51	7	1094
PT7	692	333	51	7	1083
PT8	677	330	50	8	1065

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	1125	---	---
PT2	1139	16	30
PT3	1100	53	28
PT4	1038	122	35
PT5	1122	59	56
PT6	1094	81	50
PT7	1083	90	48
PT8	1065	109	49

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de identificações corretas por faixa, usando **aresta dupla apenas quando aresta simples não criar caminho do nó origem até o nó destino, com a penalização de 10% da pontuação da sequência para cada aresta dupla do caminho**, com a sequência candidata esperada na primeira posição				
	<1,5	Entre [1,5 e 2)	Entre [2 e 2,5)	≥2,5	Total
PT1	718	337	53	7	1115
PT2	718	343	52	7	1120
PT3	696	331	49	7	1083
PT4	671	310	45	6	1032
PT5	707	340	51	8	1106
PT6	694	334	48	6	1082
PT7	692	329	48	6	1075
PT8	679	323	47	7	1056

Terceiro conjunto de espectros / Picos evidência usados na pontuação	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram considerando diferentes evidências, tendo como base o primeiro resultado		
	Total	Saíram	Entraram
PT1	1115	---	---
PT2	1120	14	19
PT3	1083	48	16
PT4	1032	106	23
PT5	1106	53	44
PT6	1082	71	38
PT7	1075	77	37
PT8	1056	95	36

Os resultados da Tabela 13 para o segundo conjunto de espectros tem o mesmo perfil dos resultados do primeiro conjunto, corroborando com o que já foi concluído. Os resultados para o terceiro conjunto de espectros, lembrando que ele é oriundo da corrida completa de uma amostra, mostra um comportamento levemente diferente dos outros dois, evidenciando a pouca identificação em *scores* acima de 2,5. Enquanto os dois primeiros conjuntos mostram que usar somente os picos do tipo y na pontuação apresenta melhores resultados, predomina levemente o uso dos picos y e b para pontuar cada nó. Podemos observar também que houve uma variação dentro de cada faixa, exceto a segunda faixa, entre 1,5 e 2, que mostra uma robustez no uso dos picos y e b na pontuação. Nas outras três faixas de *score* ocorreram variações na utilização de diferentes tipos de picos na pontuação.

Tomando como base a primeira, segunda e terceira montagens de grafos usadas, quais sejam, usando arestas simples, arestas duplas e uso de arestas duplas quando não houver caminho do nó origem até o nó destino usando arestas simples, respectivamente, e as pontuações dos nós, ou custo, usando somente a intensidade de **y** e a intensidade de **y** mais a penalização pós busca, comparou-se as diferentes listas das sequências candidatas para avaliar as entradas e saídas no topo da lista das sequências corretas entre os resultados das tabelas anteriores. As Tabelas 15, 16 e 17 mostram os totais de sequências corretas encontradas pelo algoritmo usando diferentes montagens de grafo e custos da busca para cada sequência (Y e Y+penalidade), assim como a quantidade de sequências corretas que saíram e entraram no topo da lista, nos três diferentes conjuntos de espectros.

Tabela 15: Comparação dos peptídeos que saíram e entraram do topo da classificação entre diferentes listas de sequenciamentos *de novo* corretos em diferentes montagens de grafo, **usando o primeiro conjunto de espectros**

Primeiro conjunto de testes	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram no topo da lista, tendo como base o resultado do Grafo 1+ Custo (Y)		
	Total de sequencias Corretas	Saíram	Entraram
Lista de candidatos da montagem de grafo 1 como base de comparação			
Grafo 1+ Custo (Y)	8008	---	---
Grafo 2+ Custo (Y)	10182	1920	4094
Grafo 3+ Custo (Y)	11905	0	3897
Grafo 2+ Custo (Y + penalidade)	13118	93	5203
Grafo 3+ Custo (Y) + penalidade)	13073	**0**	**5065**

Tabela 16: Comparação dos peptídeos que saíram e entraram do topo da classificação entre diferentes listas de sequenciamentos *de novo* corretos em diferentes montagens de grafo, **usando o segundo conjunto de espectros**

Segundo conjunto de testes	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram no topo da lista, tendo como base o resultado do Grafo 1+ Custo (Y)		
	Total de sequencias Corretas	Saíram	Entraram
Lista de candidatos da montagem de grafo 1 como base de comparação			
Grafo 1+ Custo (Y)	569	---	---
Grafo 2+ Custo (Y)	771	137	339
Grafo 3+ Custo (Y)	894	0	325
Grafo 2+ Custo (Y + penalidade)	976	6	413
Grafo 3+ Custo (Y) + penalidade)	978	**0**	**409**

Tabela 17: Comparação dos peptídeos que saíram e entraram do topo da classificação entre diferentes listas de sequenciamentos *de novo* corretos em diferentes montagens de grafo, **usando o terceiro conjunto de espectros**

Terceiro conjunto de testes	Quantidade de sequências corretas que saíram da primeira posição da lista de candidatos e as que entraram no topo da lista, tendo como base o resultado do Grafo 1+ Custo (Y)		
	Total de sequencias Corretas	Saíram	Entraram
Lista de candidatos da montagem de grafo 1 como base de comparação			
Grafo 1+ Custo (Y)	586	---	---
Grafo 2+ Custo (Y)	695	249	358
Grafo 3+ Custo (Y)	1001	13	428
Grafo 2+ Custo (Y + penalidade)	1125	18	557
Grafo 3+ Custo (Y) + penalidade)	1115	**13**	**542**

A comparação dos resultados com a montagem do grafo1 e pontuação Y é fundamentada no fato de que os espectros que levaram ao sequenciamento *de novo* correto são de boa qualidade, pois contém todos os fragmentos. Qualquer heurística não poderia perder estas identificações que seriam as mais seguras do grupo. Por este motivo a montagem do grafo 3 é mais adequada privilegiando os espectros que têm trajetos completos. A montagem do grafo 3 aliada à penalização obteve melhor resultado pois alia ganhos altos sem perdas. A pontuação final igual a soma das intensidades dos picos **y,** penalizada se houver aresta dupla, foi a que apresentou melhores resultados pois aliou ganhos altos sem perdas, mesmo considerando a pouca representatividade do último conjunto de espectros, pela qualidade das identificações do Comet com poucas identificações de *score* alto, acima de 2.

De uma forma geral, nos três grupos de testes, usar somente a intensidade dos picos **y** sem incluir as intensidades das evidências na pontuação dos nós do grafo mostrou-se muito eficiente, tendo tido uma leve desvantagem no terceiro conjunto de espectros com janela de 2 Th para as pontuações específicas escolhidas neste trabalho. Vale lembrar que há uma considerável quantidade de possibilidades de escolhas para este tipo de pontuação. No primeiro conjunto de testes os melhores resultados foram referentes à pontuação usando apenas a série **y**. Já no segundo e no terceiro conjunto há certa influência dos tipos de fragmentos **b**, **a**, H_2O e NH_3, o que não cria grande distância no resultado final usando apenas a série **y** como pontuação.

Portanto, os testes demonstraram que usar uma heurística de pontuação simples, neste caso utilizando apenas os picos **y**, não implica em perdas grandes no

sequenciamento *de novo*. Melhorando a aquisição dos espectros e a qualidade dos MS^2, que vem acontecendo com a evolução da espectrometria de massas, espera-se que uma pontuação simples seja capaz de sequenciar o peptídeo fragmentado na janela usando o conceito de sequenciamento *de novo*. Essa melhor qualidade dos espectros foi referenciada e comprovada por CHI *et* al [33], que diz que os espectros HCD melhoraram muito, sendo essencial para um bom sequenciamento *de novo* de peptídeos. Assim, a opção adotada para o programa DNbuilder foi a montagem do grafo 3 e a pontuação mais simples, usando somente a intensidade de **y,** com penalização de 10% pós busca das sequências com arestas duplas, se houver; exatamente como explicado na seção 5 que trata do sequenciamento *de novo* DNbuilder desenvolvido.

7 RESULTADOS

Todos os resultados apresentados neste capítulo usam a versão do programa DNbuilder desenvolvido como apresentado no Capítulo 5.

7.1 Sequenciamento *de novo* NDbuilder, Peaks, pNovo e Novor

7.1.1 Com todos os espectros de cada conjunto usado

Importante comparar o resultado do algoritmo para sequenciamento *de novo* aqui desenvolvido, DNbuilder, com os algoritmos disponíveis na literatura e já citados anteriormente, Peaks, pNovo e Novor. A Tabela 18 mostra a quantidade de peptídeos sequenciados corretamente pelos programas DNbuilder, Peaks, pNovo e Novor, usando todos os espectros do primeiro, do segundo e do terceiro conjunto. Considerou-se sequenciamento *de novo* correto quando uma sequência candidata mais ao topo de uma lista de um espectro for igual à identificação do Comet de melhor *score* para o referido espectro.

A escolha de uma tolerância de no máximo 0,02 Da no sequenciamento *de novo* é indicada quando se usa espectros registrados a partir da análise de peptídeos feitas em equipamentos de alta resolução, como Orbitrap. Quando a análise de peptídeos é realizada em equipamentos de resolução mais baixa, IonTrap, a tolerância indicada é 0,5 Da. A tolerância depende do equipamento usado devido a influência nos dados gerados. A tolerância usada neste trabalho foi de 0,02 Da, pois os resultados obtidos pelo programa DNbuilder usando a tolerância 0,5 Da não se mostraram bons para os espectros usados nos testes.

Os parâmetros utilizados na execução dos programas Peaks, pNovo e Novor foram os mesmos, quais sejam, tolerância no erro na fragmentação de 0,02 Da, peptídeos tripsinados, fragmentação CID ou HCD, modificação pós-traducional carbamidometil da cisteína, fixa, e oxidação da metionina, variável. Os programas Peaks e Novor usam, além dos parâmetros citados, uma tolerância no erro do precursor, onde o valor de 15 ppm foi usado nos testes.

Tabela 18: Quantidade de peptídeos sequenciados corretamente na primeira posição encontrados pelos programas DNbuilder, Peaks, pNovo e Novor para os três conjuntos de espectros. Em cinza estão marcados os melhores resultados para cada conjunto de espectros

Conjunto de Espectros	Quantidade de peptídeos sequenciados *de novo* com a sequência candidata correta na primeira posição			
	DNbuilder	Peaks	pNovo	Novor
Primeiro (%)	13073 (21,0%)	10396 (16,7%)	6338 (10,2%)	9209 (14,8%)
Segundo (%)	978 (19,3%)	814 (16,1%)	800 (15,8%)	535 (10,6%)
Terceiro (%)	1115 (6,7%)	1267 (7,6)	712 (4,3%)	862 (5,2%)

O programa DNbuilder mostrou-se eficiente nos dois primeiros conjuntos de espectros. DNbuilder conseguiu sequenciar corretamente 13073 peptídeos, representando 21,0% do total, seguido pelo programa Peaks, 10396, 16,7%. Novor sequenciou 9209, 14,8%, enquanto que pNovo sequenciou corretamente 6338 peptídeos, 10,2% do total. Cabe aqui esclarecer que sequenciar *de novo* corretamente um espectro é ter a melhor sequência candidata igual a sequência de melhor *score* identificada pelo Comet para o referido espectro.

O programa DNbuilder também se mostrou eficiente no sequenciamento *de novo* do segundo conjunto, onde conseguiu sequenciar corretamente 978 peptídeos, que representa 19,3% do total de espectros usados, os 535 sequenciados corretamente pelo programa Novor representam 10,6%, 653 pelo Peaks representam 12,9% e os 800 do programa pNovo representam 15,8% dos peptídeos sequenciados corretamente.

Já no terceiro conjunto de espectros, o programa Peaks foi mais eficiente, sequenciando corretamente 1267 espectros, 7,6% do total, enquanto DNbuilder conseguiu sequenciar 1115 espectros, 6,7%. Os programas Novor e pNovo conseguiram sequenciar corretamente 862, 5,2%, e 712, 4,3%, peptídeos, respectivamente.

A razão pela qual DNbuilder apresentou tanto sucesso é porque considerou arestas duplas. Se somente arestas simples fossem consideradas, resultados apresentados na seção 6.3, os resultados seriam 8008 (12,9%), 569 (11,2%) e 586 (3,53%), respectivamente. Neste caso o algoritmo perderia para quase todos os outros. Importante considerar aqui os tipos de fragmentos adotados no sequenciamento *de novo*. Os programas adotam comumente os fragmentos do tipo **b**. Não houve estudo aqui de quantos espectros teriam os fragmentos **b** completos e isso influenciaria muito os resultados. Parece que há mais espectros com fragmentos **b** completos que **y** nestes

conjuntos, o que faria sentido, pois foram os espectros usados nos testes de algoritmos que usam a série **b** para sequenciar e devem ter sido escolhidos especialmente para os testes já que não haveria como sequenciar se os fragmentos não estivessem presentes de forma que a maioria destes espectros deve conter mesmo mais fragmentos tipo **b** que **y**, como mostram os resultados. Não temos, contudo, a informação de quantos espectros destes conjuntos tem a sequência dos fragmentos tipo **b** completa afim de fazer uma análise comparativa dos resultados.

Nota-se que o resultado para o DNbuilder considerando o primeiro conjunto de espectros chegou muito próximo da metade do rendimento máximo possível de 44,23% de espectros sequenciáveis, determinado na seção 6.3; 21% foram sequenciados pelo DNbuilder. O mesmo acontecendo com o segundo conjunto, onde o máximo de sucesso possível para o segundo conjunto, de 38,98%, e o programa DNbuilder ficou muito próximo de alcançar a metade, chegando a 19,3%. Seria preciso analisar melhor estes espectros e verificar se haveria alguma correlação com o tipo de fragmentação CID/HCD. Talvez haja a necessidade de sofisticar o sistema de pontuação para melhorar a busca, mas não se trata de tarefa simplificada. Sabemos que a maioria dos algoritmos leva em conta treinamentos de um volume considerável de espectros anotados para melhorar a interpretação e desempenho dos algoritmos, mas a tendência de se adquirir cada vez melhores espectros deve derrubar esta necessidade. Outra opção viável seria tentar sequenciar usando a serie **b** se não houver resultado satisfatório para a série **y**. Estas opções comporiam idéias para um trabalho futuro. De qualquer forma, considerar a aresta dupla melhorou consideravelmente o resultado do sequenciamento *de novo*, mostrando confiabilidade.

Ao observar o bom desempenho do programa DNbuilder usando arestas duplas penalizadas com busca pela série **y**, enquanto os programas usados, com exceção do Peaks, deixam claro que usam íons da série **b**, surge a questão de se saber quantos peptídeos foram sequenciados em comum e quantos foram sequenciados mutuamente exclusivos. As Tabelas 19, 20 e 21 mostram os sequenciamentos *de novo* corretos de cada programa confrontados uns com os outros, para o primeiro, segundo e terceiro conjunto de espectros, respectivamente.

Cabe aqui esclarecer que o tempo gasto por cada programa na execução do sequenciamento *de novo* de cada conjunto de espectros, apesar de importante, não foi

considerado nos testes. O que de fato importou foi a contagem dos de acertos de cada programa para confrontar as diferentes metodologias.

Tabela 19: Número de acertos comuns e exclusivos realizados pelos programas DNbuider, Peaks, pNovo e Novor para o primeiro conjunto de espectros

Programa que obteve acerto	Número de acertos
Só DNbuilder	6406
Só Peaks	1249
Só Novor	846
Só pNovo	444
Acertos comuns a todos os programas	3442
Resultados comuns entre dois programas	
DNbuilder X Peaks	465
DNbuilder X Novor	356
DNbuilder X pNovo	132
Peaks X Novor	1787
Peaks X pNovo	575
Novor X pNovo	203
Resultados comuns entre três programas	
DNbuilder X Peaks X Novor	1547
DNbuilder X Peaks X pNovo	514
DNbuilder X Novor X pNovo	211
Peaks X Novor X pNovo	817

Tabela 20: Número de acertos comuns e exclusivos realizados pelos programas DNbuider, Peaks, pNovo e Novor para o segundo conjunto de espectros

Programa que obteve acerto	Número de acertos
Só DNbuilder	235
Só Peaks	142
Só Novor	49
Só pNovo	132
Acertos comuns a todos os programas	196
Resultados comuns entre dois programas	
DNbuilder X Peaks	71
DNbuilder X Novor	18
DNbuilder X pNovo	259
Peaks X Novor	104
Peaks X pNovo	75
Novor X pNovo	10
Resultados comuns entre três programas	
DNbuilder X Peaks X Novor	109
DNbuilder X Peaks X pNovo	79
DNbuilder X Novor X pNovo	11
Peaks X Novor X pNovo	38

Tabela 21: Número de acertos comuns e exclusivos realizados pelos programas DNbuider, Peaks, pNovo e Novor para o terceiro conjunto de espectros

Programa que obteve acerto	Número de acertos
Só DNbuilder	537
Só Peaks	334
Só Novor	78
Só pNovo	54
Acertos comuns a todos os programas	259
Resultados comuns entre dois programas	
DNbuilder X Peaks	55
DNbuilder X Novor	25
DNbuilder X pNovo	28
Peaks X Novor	217
Peaks X pNovo	84
Novor X pNovo	26
Resultados comuns entre três programas	
DNbuilder X Peaks X Novor	88
DNbuilder X Peaks X pNovo	92
DNbuilder X Novor X pNovo	31
Peaks X Novor X pNovo	138

Os mesmos resultados mostrados nas tabelas acima também foram representados em Diagramas de Venny como alternativa de visualização. A Figura 17 mostra os acertos exclusivos e comuns entre os diferentes programas em Diagramas de Venny. Os autores do programa usado para a criação dos diagramas relatam que não há um artigo científico associado ao programa. Esse programa pode ser encontrado no sítio http://bioinformatics.psb.ugent.be/webtools/Venn/.

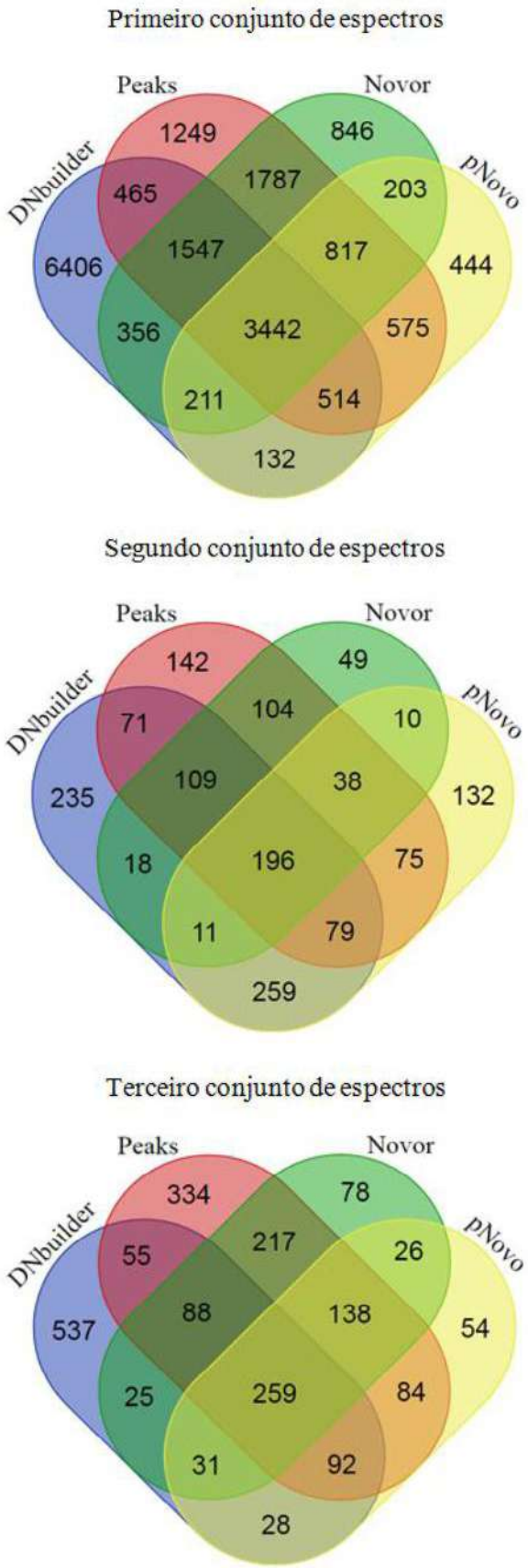


Figura 17: Diagrama de Venny para o número de acertos comuns e exclusivos realizados pelos programas DNbuider, Peaks, pNovo e Novor para os três conjuntos de espectros usados

Um número que chama a atenção é a quantidade de acertos exclusivos do programa DNbuilder no primeiro conjunto de espectros, 6406, de um total de 13073. O mesmo aconteceu no terceiro conjunto de espectros, onde o maior número de acertos foi conseguido exclusivamente pelo DNbuilder. Já no segundo conjunto, exclusivamente DNbuilder ficou em segundo lugar, com 235 acertos, perdendo apenas para o conjunto DNbuilder X pNovo, com 259 acertos.

O resultado alcançado exclusivamente pelo DNbuilder pode ser em consequência da escolha do uso da série **y** na criação do grafo. Esse fato também alimenta uma possibilidade futura do uso integrado das buscas no grafo usando também a série **b** para aumentar quantidade de sequenciamentos *de novo* pelo programa.

7.1.2 Combinações de aminoácidos para os programas Peaks e pNovo

Os programas Peaks e pNovo retornam sequências com pontuações iguais para muitos espectros. O mesmo fato é observado nos resultados do programa DNbuilder em razão do uso de arestas duplas. Acontece que na contabilização dos acertos, DNbuilder considera mais de uma sequência no topo da lista em decorrência de diferentes ordens de aminoácidos das arestas duplas, com pontuações iguais. Assim, foram contabilizados como acertos dos programas Peaks e pNovo, qualquer das sequências de mesma pontuação que estejam no topo da lista do programa. Exemplificando: a sequência candidata mais ao topo da lista pelo programa Peaks ou pNovo para um espectro é GGGNR, de pontuação 16,7. A segunda sequência candidata da lista é GGNGR e possui a mesma pontuação. A sequência esperada é GGNGR. Nesse caso a segunda sequência candidata que se encontra na segunda posição da lista é considerada como um acerto.

Cabe salientar que o programa Novor retorna apenas uma solução, assim, não há possibilidade de mais de uma sequência com a mesma pontuação.

A Tabela 22 mostra os sequenciamentos *de novo* corretos para os três conjuntos de espectros, agora considerando como correto qualquer das sequências com a mesma pontuação no topo da lista que sejam iguais a identificação do Comet anotada, ou seja, de melhor *score*.

Tabela 22: Quantidade de peptídeos sequenciados corretamente na primeira posição encontrados pelos programas Peaks e pNovo para os três conjuntos de espectros, considerando diferentes ordens de aminoácidos para sequências com a mesma pontuação que a melhor sequência

Conjunto de Espectros com combinação de aminoácidos	Quantidade de peptídeos sequenciados *de novo* com a sequência candidata correta na primeira posição	
	Peaks	pNovo
Primeiro (%)	11804 (18,9%)	8765 (14,1%)
Segundo (%)	887 (17,3%)	923 (18,2%)
Terceiro (%)	1308 (7,9)	816 (5,0%)

A utilização de combinação de aminoácidos nos resultados dos programas Peaks e pNovo melhoram a contagem de sequências corretas, mas ainda assim, os acertos exclusivos do programa DNbuilder foram maiores em todas as situações testadas. Os resultados foram representados em Diagramas de Venny. A Figura 18 representa os resultados comuns e exclusivos dos programas DNbuilder, Peaks, pNovo e Novor, a diferença para a Figura 17 é que a contagem dos acertos para os programas Peaks e pNovo consideram como acerto qualquer das sequências com mesma pontuação da sequência do topo da lista.

Figura 18: Diagrama de Venny para o número de acertos comuns e exclusivos realizados pelos programas DNbuider, Peaks, pNovo e Novor para os três conjuntos de espectros usados e combinação de aminoácidos nos resultados dos programas Peaks e pNovo

7.1.3 PatternLab para obter espectros de qualidade

Os testes para validar a metodologia para sequenciamento *de novo* proposta usaram todos os espectros identificados pelo programa Comet, independentemente do score e deltaCN. Acontece que muitas identificações são de qualidade duvidosa. Para dar uma maior confiança nos resultados do sequenciamento *de novo*, os mesmos testes foram realizados usando somente espectros de boa qualidade. A qualidade de cada espectro dos três conjuntos foi determinada pelo programa PatternLab [55]. Os parâmetros usados foram os *default*, exceto, *Delta Mass PPM*=20, e *Primay Score*=2,0. A Tabela 23 mostra os sequenciamentos *de novo* corretos para os dois primeiros conjuntos de espectros, selecionados pelo filtro de qualidade PattenLab. Os acertos dos programas Peaks e pNovo consideram os acertos das sequências candidatas que possuem a mesma pontuação da melhor sequência candidata, mas com inversão de aminoácidos.

Tabela 23: Quantidade de peptídeos sequenciados corretamente na primeira posição encontrados pelos programas DNbuilder, Peaks, pNovo e Novor para os espectros selecionados pelo filtro PatternLab. Em cinza estão marcados os melhores resultados para cada conjunto de espectros

Conjunto de Espectros selecionados pelo filtro PatternLab	Quantidade de peptídeos sequenciados *de novo* com a sequência candidata correta na primeira posição			
	DNbuilder	Peaks	Novor	pNovo
Primeiro – 32964 espectros (%)	11201 (33,9%)	10217 (31,0%)	8751 (26,5%)	8186 (24,8%)
Segundo – 2366 espectros (%)	812 (34,3%)	691 (29,4%)	458 (19,4%)	684 (28,9%)

Os resultados com os espectros selecionados PatternLab confirmam que o programa DNbuilder encontrou mais sequências corretas na primeira posição, seguido pelo Peaks. O terceiro conjunto não foi representado na tabela, pois nenhum espectro foi selecionado pelo PatternLab, confirmando a qualidade duvidosa das identificações.

7.2 Sequenciamento *de novo* usando espectros multiplex simulados

Os espectros multiplex DDA (*Data Dependent Acquisition*) são obtidos a partir das análises realizadas por espectrômetros de massa, de acordo com a definição pelo operador do equipamento para o tamanho da janela de isolamento dos peptídeos precursores. Os precursores selecionados são fragmentados e seus íons são registrados no espectro. Se janelas amplas forem selecionadas como padrão da análise, todos os

peptídeos contidos na janela, e não somente o precursor, serão fragmentados juntos gerando espectros ditos multiplex.

Para estudar os padrões que permitam desenvolver uma heurística para sequenciamento *de novo* dos peptídeos cujos fragmentos estejam presentes nos espectros multiplex, optou-se por simular espectros multiplex misturando espectros MS^2 de peptídeos isolados em janelas de 2 Th. Estes espectros foram aqui denominados de espectros multiplex simulados. Ao todo foram usados cinco espectros HCD anotados, combinados dois a dois e três a três, gerando 10 espectros multiplex simulados contendo fragmentos de 2 peptídeos e 10 espectros multiplex simulados contendo fragmentos de 3 peptídeos.

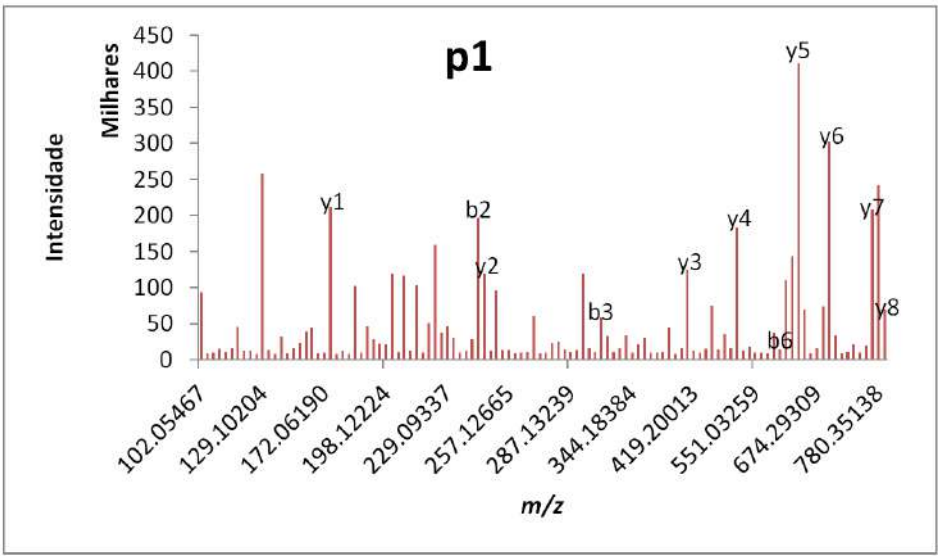
Os espectros foram escolhidos aleatoriamente entre os espectros MS^2 do primeiro conjunto de espectros, mas somente dentre os espectros identificados com *score* Comet acima de 2,5. Esses espectros foram aqui designados como p1, p2, p3, p4 e p5. Os espectros multiplex foram gerados na seguinte ordem de combinação: E12=p1+p2; E13=p1+p3; E14=p1+p4; .... ; E45=p4+p5. A ordem dos peptídeos que aparecem nas Tabelas respeita sempre a ordem da soma: o primeiro peptídeo sequenciado de E14, por exemplo, é p1, ou p_1^1, e o segundo, p4, ou p_4^2, sempre na ordem em que aparecem na soma acima. O primeiro peptídeo do espectro simulado E_{ij} , p_i^1, tem sempre a média das intensidades dos picos maior que a do segundo p_j^2.

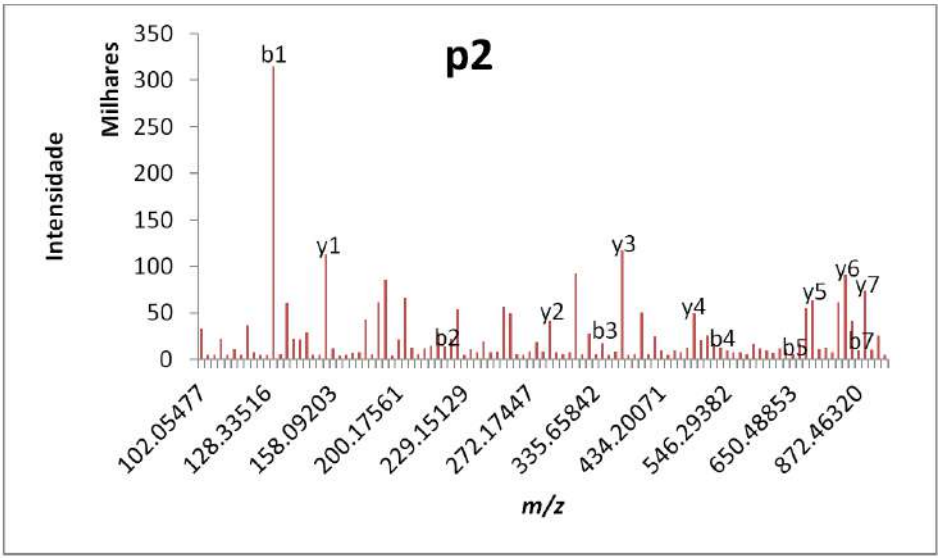
A Tabela 24 mostra as massas dos peptídeos precursores selecionados, sua *m/z*, o número de picos do espectro, a média aritmética dos picos dos espectros desses peptídeos e o *DeltaCN*, que é a diferença absoluta do *score* da primeira sequência candidata do Comet e o *score* da segunda sequência candidata do Comet do mesmo precursor, ponderado pelo primeiro *score*, $DeltaCN = \Delta score_{2,1}/score_1$.

Tabela 24: *m/z*, número de picos e média aritmética dos picos dos espectros HCD selecionados

Precursor	*m/z*(carga 2)	Número de picos	Média das intensidades de todos os picos do espectro	DeltaCN (normalizado pelo maior *score*)
p1	1023,44948	112	47908,49688	0,2982083
p2	1018,55752	105	24869,82752	0,1645656
p3	1030,61538	90	22527,99756	0,2451711
P4	1040,54874	65	15369,94662	0,3586237
p5	1044,53251	123	2001,129106	0,1764156

Os picos dos espectros foram plotados em gráficos para melhor visualização das intensidades de cada espectro. A Figura 19 mostra os espectros originais de p1, p2, p3, p4 e p5.





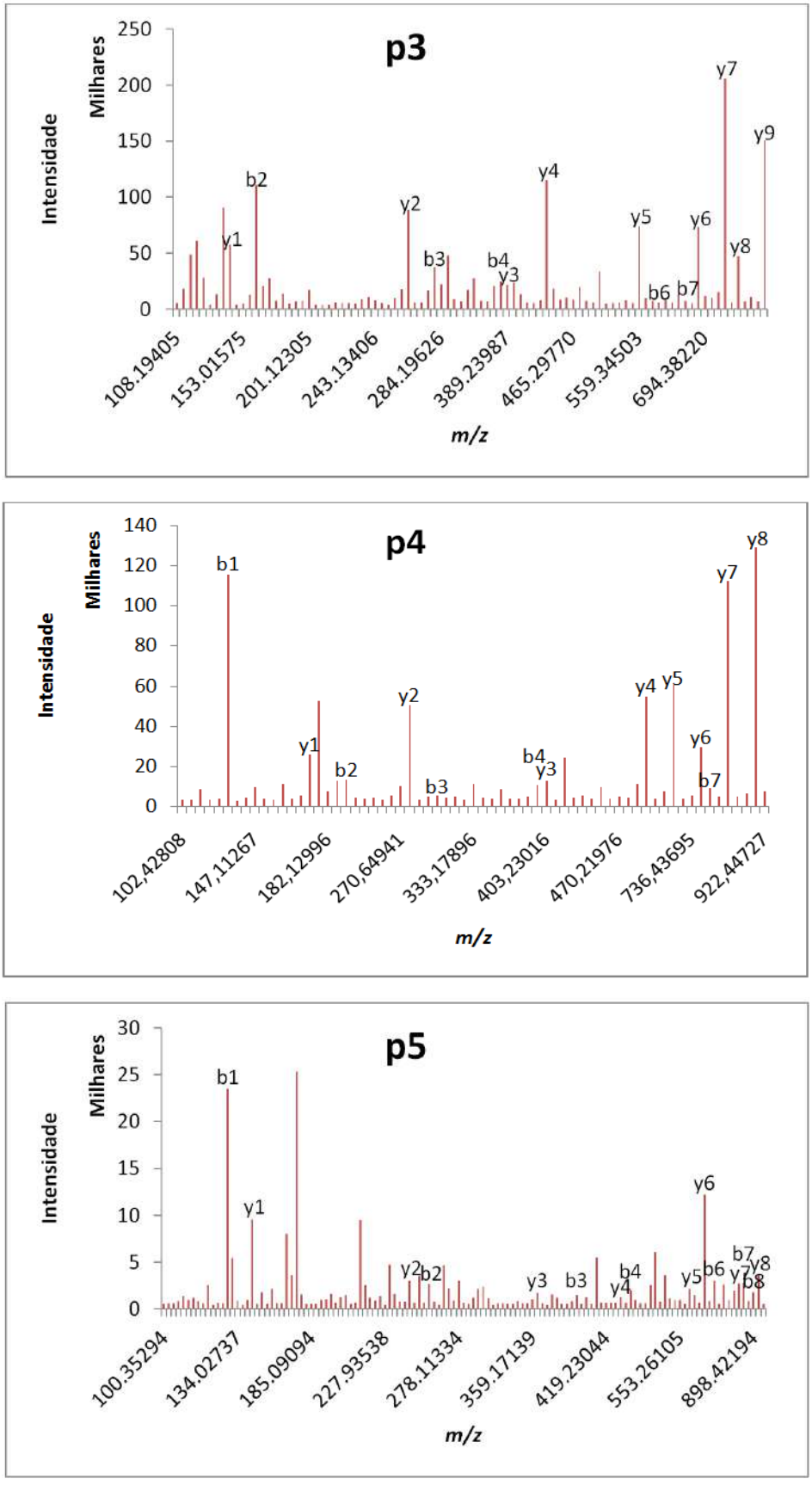


Figura 19: Espectros de MS^2 dos peptídeos p1, p2, p3, p4 e p5

Como os espectros escolhidos são de boa qualidade, o programa DNBuilder, usando tolerância igual a 0,02 Da na construção dos grafos, sequencia corretamente os peptídeos esperados destes espectros na primeira posição na sua lista de sequências candidatas. Usando a mesma tolerância, os programas Peaks, pNovo e Novor foram executados, quase todos identificando os peptídeos corretamente como mostrado na Tabela 25.

Tabela 25: Posição da sequência esperada dentro das listas de candidatos geradas pelos programas DNBuilder, Peaks, pNovo e Novor

Espectro	Posição na lista de sequências candidatas			
	DNbuilder	Peaks	pNovo	Novor
Tolerância 0,02				
p1	1	1	0	1
p2	1	1	1	1
p3	1	1	1	1
p4	1	1	1	1
p5	1	1	1	1

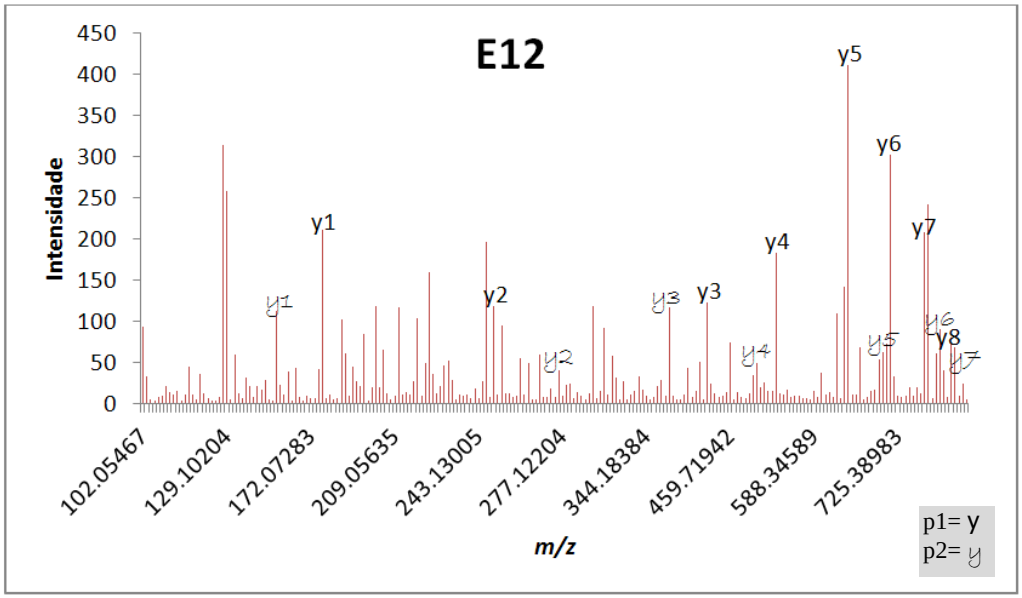
A Tabela 26 mostra número de picos dos espectros multiplex simulados, o número de nós e o número de arestas dos grafos criados pelo DNBuilder para cada peptídeo alvo, usando tolerância 0,02 Da.

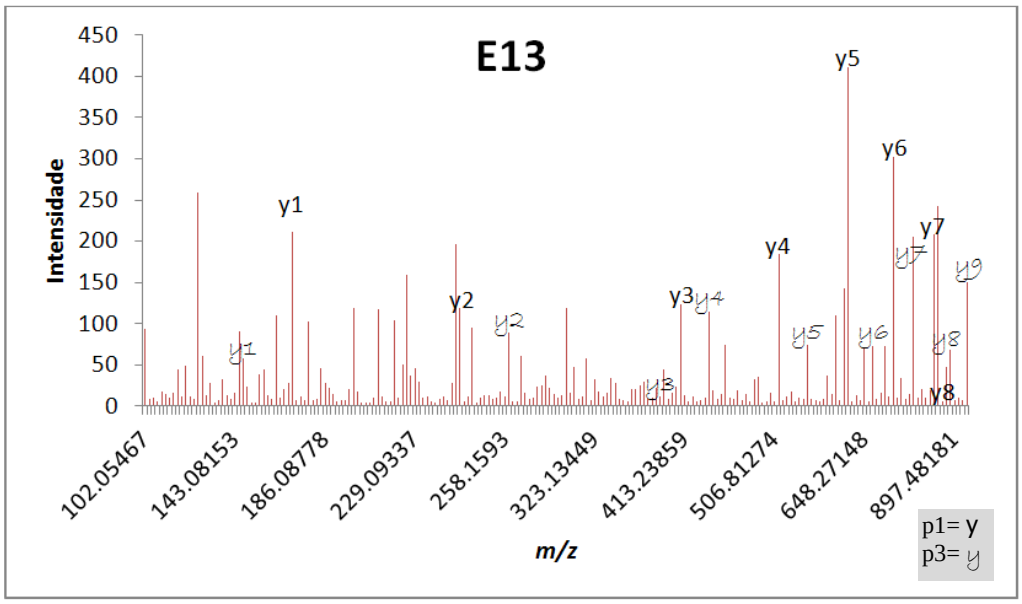
Tabela 26: Número de picos dos espectros multiplex simulados, número de nós e arestas de cada grafo montado para cada peptídeo alvo, usando a tolerância 0,02 Da

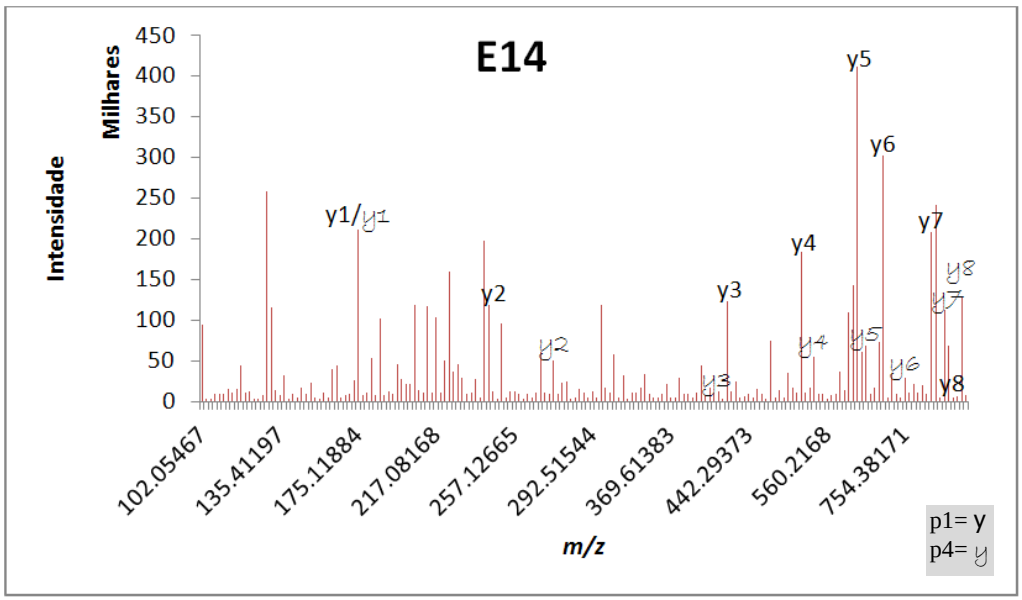
Espectro E*ij*	Número de picos do espectro	Grafos por peptídeo			
		Número de nós		Número de arestas	
		p_i^1	p_j^2	p_i^1	p_j^2
Aresta com tamanho de 1 até dois aminoácidos					
E12	217	79	444	73	351
E13	202	80	436	44	146
E14	177	63	299	50	150
E15	235	92	576	87	471
E23	195	91	549	66	303
E24	170	63	270	71	310
E25	228	75	362	98	548
E34	155	43	139	68	311
E35	213	56	195	103	598
E45	188	74	323	72	291

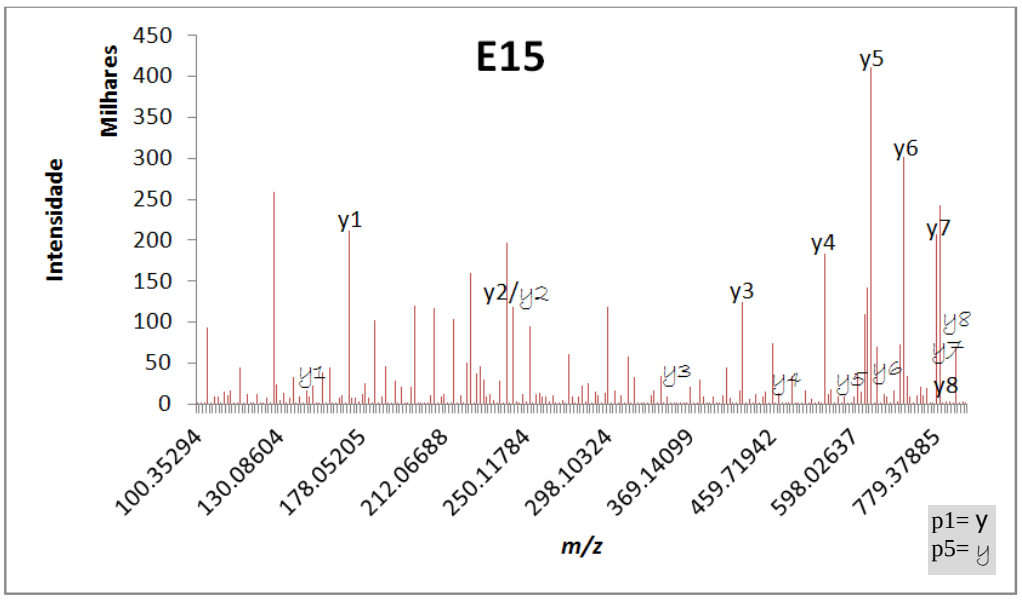
Como explicado anteriormente, podemos observar na Tabela 26 a união dos picos de dois em dois espectros. O espectro E12, com 217 picos, é o resultado da união dos espectros p1 e p2. Construiu-se dois diferentes grafos usando as massas tanto de p1 quanto a massa de p2. O grafo criado para E12 usando a massa de p1 possui 79 nós e 444 arestas. O grafo criado para E12 usando a massa de p2 possui 73 nós e 351 arestas. Os outros espectros multiplex simulados seguiram o mesmo procedimento.

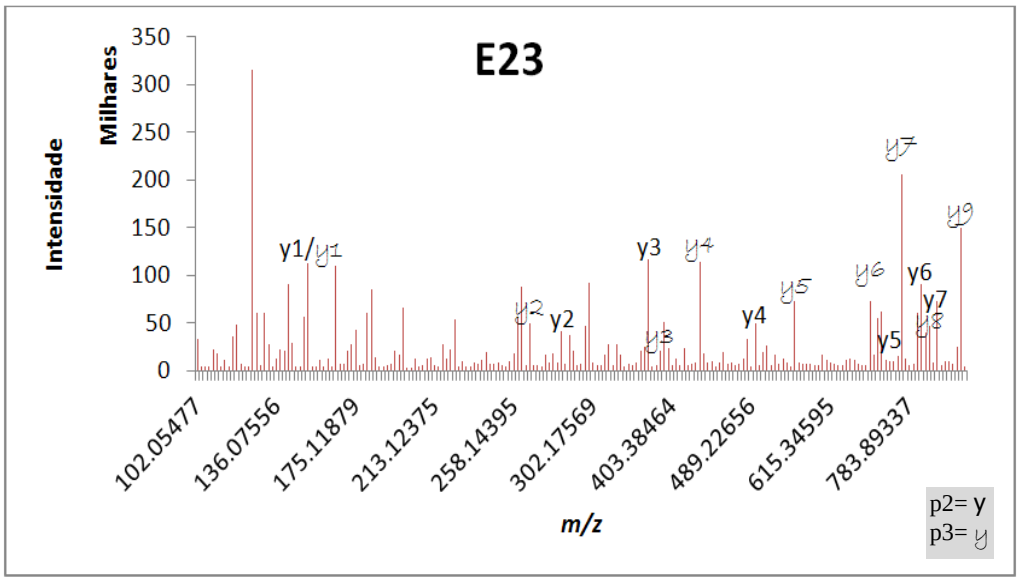
Igualmente ao que foi feito com os cinco espectros originais, os espectros simulados foram plotados, mostrados na Figura 20. Os picos das séries **y** dos dois peptídeos foram marcados usando tipos diferentes de fonte de caracteres para facilitar a visualização. Diferentemente dos espectros p1, p2, p3, p4 e p5, os espectros combinados não possuem a marcação dos picos da série **b** por causa da poluição visual causada no gráfico devido à baixa intensidade de muitos picos dessa série.

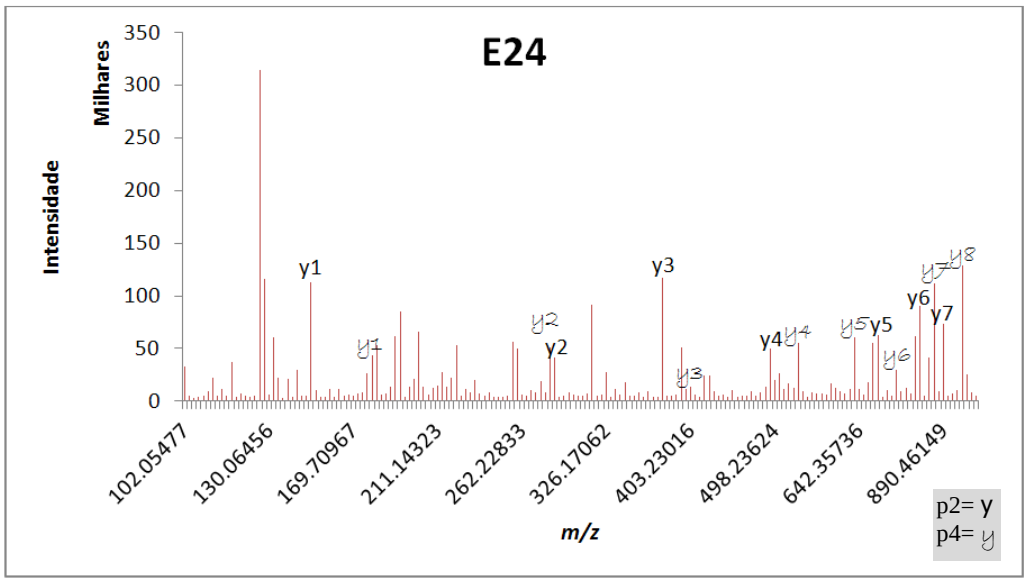


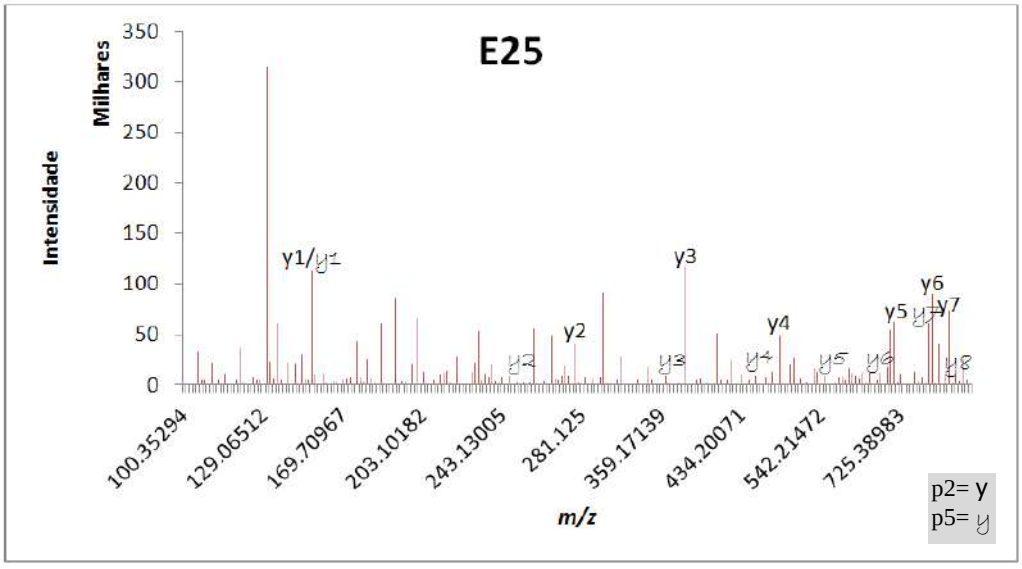

E13
Intensidade
450
400
350
300
250
200
150
100
50
0
y1
y2
y3
y4
y5
y6
y7
y8
102.05467
143.08153
186.08778
229.09337
258.1593
323.13449
413.23859
506.81274
648.27148
897.48181
m/z
p1= y
p3=

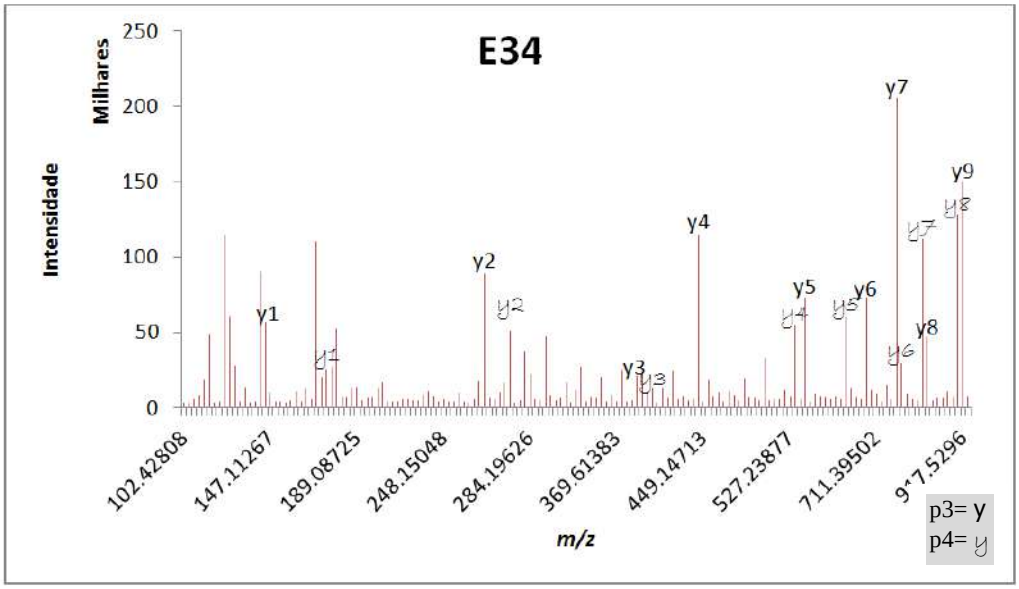

E14
Intensidade
Milhares
450
400
350
300
250
200
150
100
50
0
y1/
y2
y3
y4
y5
y6
y7
y8
102.05467
135.41197
175.11884
217.08168
257.12665
292.51544
369.61383
442.29373
560.2168
754.38171
m/z
p1= y
p4=

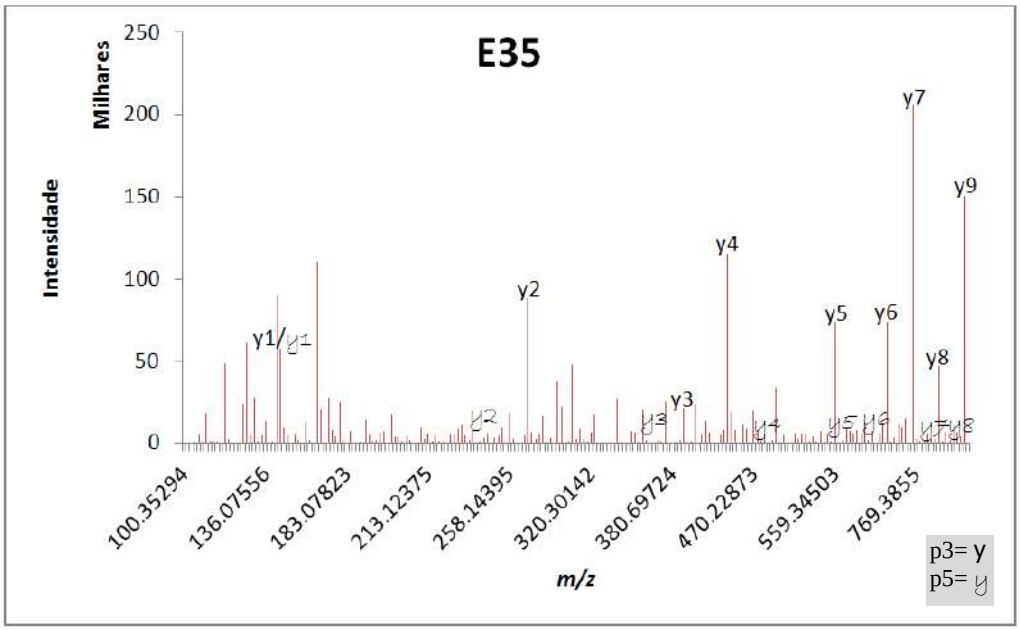

E15
Intensidade
Milhares
450
400
350
300
250
200
150
100
50
0
y1
y2/
y3
y4
y5
y6
y7
y8
100.35294
130.08604
178.05205
212.06688
250.11784
298.10324
369.14099
459.71942
598.02637
779.37885
m/z
p1= y
p5=


E23
Intensidade
Milhares
350
300
250
200
150
100
50
0
y1/y1
y2
y3
y4
y5
y6
y7
y2
y3
y4
y5
y6
y7
y8
y9
102.05477
136.07556
175.11879
213.12375
258.14395
302.17569
403.38464
489.22656
615.34595
783.89337
m/z
p2= y
p3= y


E24
Intensidade
Milhares
350
300
250
200
150
100
50
0
y1
y2
y3
y4
y5
y6
y7
y1
y2
y3
y4
y5
y6
y7
y8
102.05477
130.06456
169.70967
211.14323
262.22833
326.17062
403.23016
498.23624
642.35736
890.46149
m/z
p2= y
p4= y


E25
Intensidade
Milhares
350
300
250
200
150
100
50
0
y1/y1
y2
y3
y4
y5
y6
y7
y2
y3
y4
y5
y6
y7
y8
100.35294
129.06512
169.70967
203.10182
243.13005
281.125
359.17139
434.20071
542.21472
725.38983
m/z
p2= y
p5= y


E34
Milhares
Intensidade
250
200
150
100
50
0
y1
y2
y3
y4
y5
y6
y7
y8
y9
102.42808
147.11267
189.08725
248.15048
284.19626
369.61383
449.14713
527.23877
711.39502
m/z
p3= y
p4= y


E35
Milhares
Intensidade
250
200
150
100
50
0
y1/y1
y2
y3
y4
y5
y6
y7
y8
y9
100.35294
136.07556
183.07823
213.12375
258.14395
320.30142
380.69724
470.22873
559.34503
769.3855
m/z
p3= y
p5= y

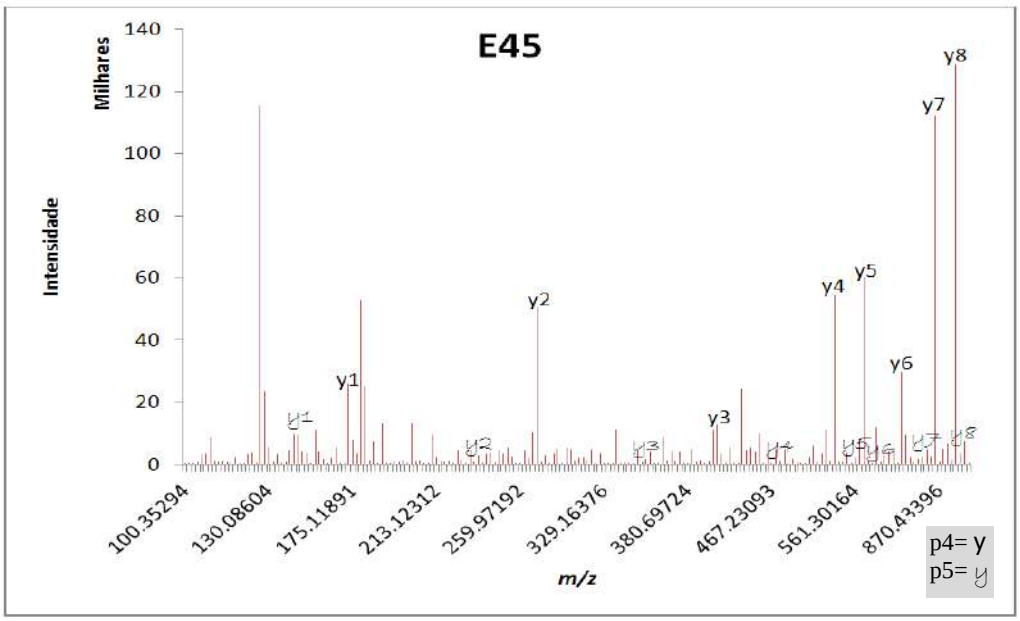


Figura 20: Gráficos das intensidades dos espectros multiplex simulados. Os picos da série **y** dos dois peptídeos estão marcados usando diferentes fonte de caracteres.

O objetivo ao executar um programa para sequenciamento *de novo* em espectros multiplex é que se consiga buscar as duas sequências corretas no topo da lista. Dessa forma, usando os espectros multiplex simulados, executou-se a versão do programa DNbuilder que constrói grafo usando arestas duplas penalizadas para todo o espectro caso um grafo não seja criado usando arestas simples. A tolerância usada foi 0,02 Da. Os resultados são apresentados na Tabela 27, que mostra a identificação do espectro simulado (anotação do espectro), o sequenciamento *de novo* do peptídeo alvo pelo DNbuilder e o percentual de acerto de aminoácidos na posição correta.

Tabela 27: Sequenciamento *de novo* usando os espectros multiplex simulados, considerando cada peptídeo precursor, a sequência esperada, a sequência encontrada e o percentual de acerto dos aminoácidos

Espectro simulado	Precursor	Sequência esperada	Sequência encontrada	% de acerto de aminoácido na posição correta
E12	p1	EDAANNYAR	EDAANNYAR	100
E12	p2	KVDWLTEK	KVDWLMAR	62,5
E13	p1	EDAANNYAR	EDAANNYAR	100
E13	p3	LGLSTLGELK	LGLSTLGELK	100
E14	p1	EDAANNYAR	EDAANNYAR	100
E14	p4	KASDVHEVR	KASDVHEVR	100
E15	p1	EDAANNYAR	EDAANNYAR	100
E15	p5	KEEPSNNVK	KHAANNYAR	22,2
E23	p2	KVDWLTEK	TGLSTLGELK	12,5
E23	p3	LGLSTLGELK	LGLSTLGELK	100
E24	p2	KVDWLTEK	KVDWLTEK	100
E24	p4	KASDVHEVR	KASDVHEVR	100
E25	p2	KVDWLTEK	KVDWLTEK	100
E25	p5	KEEPSNNVK	KEEPSPTEK	66,7
E34	p3	LGLSTLGELK	LGLSTLGELK	100
E34	p4	KASDVHEVR	KASNGHGELK	44,4
E35	p3	LGLSTLGELK	LGLSTLGELK	100
E35	p5	KEEPSNNVK	FKEGHGELK	33,3
E45	p4	KASDVHEVR	KASDVHEVR	100
E45	p5	KEEPSNNVK	KFQTHEVR	22,2

Partindo do pressuposto que o sequenciamento *de novo* do peptídeo mais intenso seria corretamente sequenciado, peptídeo designado como $alvo_1$, o objetivo seria identificar o segundo peptídeo alvo, $alvo_2$. A estratégia foi primeiramente a redução dos picos dos fragmentos **y** e **b** da sequência encontrada para o $alvo_1$, mais intenso, e depois o abatimento das intensidades dos picos mais intensos do espectro. Para o abatimento considerou-se a redução da Eq. (23) aplicada sobre um percentual dos picos mais altos do espectro simulado. O fator de redução que aparece na Eq.(23) é 0,5, mas testou-se também 0,25.

Optou-se por não se remover picos do espectro referentes à sequência encontrada para o peptídeo mais intenso, mas sim a redução da intensidade dos picos de fragmentos tipo **y** e **b** da sequência do peptídeo mais intenso, porque a perda de um pico pode ser fatal para o sequenciamento *de novo* se, por exemplo, houvesse coincidência de massas de fragmentos nas duas sequências. As heurísticas testadas para modificar os espectros multiplex simulados foram:

1. H1: Alterar para 1 as intensidades dos picos **y** e **b** da sequência encontrada para o peptídeo mais intenso;

2. H2: Alterar para 1 as intensidades dos picos **y** da sequência encontrada para o peptídeo mais intenso; abater 33% dos picos mais intensos do espectro usando Eq. (22), com *fator*=0,25.;
3. H3: Alterar para 1 as intensidades dos picos **y** e **b** da sequência encontrada para o peptídeo mais intenso; abater 33% dos picos mais intensos do espectro usando Eq.(22) com *fator*=0,50;
4. H4: Alterar para 1 as intensidades dos picos **y** da sequência encontrada para o peptídeo mais intenso; e reduzir 33% dos peptídeos mais intensos do espectro usando Eq. (22) com *fator*=0,50;
5. H5: Alterar para 1 as intensidades dos picos **y** da sequência encontrada para o peptídeo mais intenso; e reduzir 33% dos peptídeos mais intensos do espectro usando Eq. (22) com *fator*=0,50, exceto para possíveis y1, reduzido apenas pela relação da intensidade do peptídeo menos intenso pela intensidade do peptídeo mais intenso, e
6. H6: Alterar para 1 as intensidades dos picos **y** e **b** da sequência encontrada para o peptídeo mais intenso; e reduzir 33% dos peptídeos mais intensos do espectro usando Eq. (22) com *fator*=0,5; exceto y1, que recebe a maior intensidade entre os picos da lisina e arginina (de massa inteira 147 ou 175). Na ausência das massas desses dois picos da série **y**, as intensidades dos seus complementos são utilizadas.

Estas heurísticas foram testadas e o resultado é apresentado na Tabela 28, mostrando uma marcação com um X nas heurísticas que acertaram no sequenciamento *de novo* dos peptídeos menos intensos. As marcações em verde indicam as sequências do segundo peptídeo que já haviam sido sequenciadas corretamente na primeira tentativa antes de aplicar as heurísticas (vide Tabela 26).

Tabela 28: Peptídeo menos intenso sequenciado corretamente usando diferentes heurísticas multiplex. Estão marcados com fundo verde os peptídeos menos intensos sequenciados corretos originalmente. A marcação com um X representa que houve um sequenciamento *de novo* correto do espectro para a heurística

Heurística	E12	E13	E14	E15	E24	E25	E34	E35	E45
H1	X	X	X		X		X		X
H2	X	X	X	X		X	X		X
H3	X	X	X	X	X	X	X		X
H4	X	X	X	X	X	X	X	X	X
H5	X	X	X	X	X	X	X		X
H6		X	X	X	X	X	X	X	X

De uma forma geral, todas as heurísticas melhoraram os resultados, mas algumas perderam sequências que originalmente estavam corretas, como H2. Duas heurísticas apenas sequenciaram corretamente os peptídeos menos intensos do espectro E35, a H4 e H6.

As três heurísticas que apresentaram melhores resultados são H3, H4 e H6, marcadas em cinza na Tabela 28. Dentre as três, a de melhor resultado neste experimento foi H4 seguida de perto por H6 e H3.

Usando os mesmos p1, p2, p3, p4 e p5, novos espectros simulados foram criados, agora combinando de 3x3 espectros. Com os espectros combinados 3x3 executou-se o programa DNbuilder usando arestas duplas penalizadas para todo o espectro apenas quando não existir grafo criado usando arestas simples. A Tabela 29 mostra os espectros simulados criados, a quantidade de picos e as sequências esperadas e encontradas sem uso de heurísticas.

Tabela 29: Sequenciamento *de novo* usando os espectros multiplex simulados 3x3, considerando cada peptídeo precursor, a sequência esperada, a sequência encontrada e o percentual de acerto dos aminoácidos. Em verde estão marcadas as sequências sequenciadas corretamente

Espectro simulado	Número de picos	Precursor	Sequência esperada	Sequência encontrada	% de acerto de aminoácido na posição correta
E123	307	p1	EDAANNYAR	EDAANNYAR	100
E123	307	p2	KVDWLTEK	TGLSTLGLTR	0
E123	307	p3	LGLSTLGELK	LGLSTLGLTR	60,0
E124	282	p1	EDAANNYAR	EDAANNYAR	100
E124	282	p2	KVDWLTEK	KVDWLMAR	62,5
E124	282	p4	KASDVHEVR	KASDVHEVR	100
E125	340	p1	EDAANNYAR	EDAANNYAR	100
E125	340	p2	KVDWLTEK	KKAANEMAR	14,3
E125	340	p5	KEEPSNNVK	KHAANNYAR	33,3
E134	267	p1	EDAANNYAR	EDAANNYAR	100
E134	267	p3	LGLSTLGELK	LGLSTLLSAR	60,0
E134	267	p4	KASDVHEVR	KASNGHLSAR	55,5
E135	325	p1	EDAANNYAR	EDAANNYAR	100
E135	325	p3	LGLSTLGELK	LGLSTLGLTR	60,0
E135	325	p5	KEEPSNNVK	KHAANNYAR	33,3
E145	300	p1	EDAANNYAR	EDAANNYAR	100
E145	300	p4	KASDVHEVR	KASDVHEVR	100
E145	300	p5	KEEPSNNVK	KHAANNYAR	33,3
E234	260	p2	KVDWLTEK	TGLSTLGELK	14,3
E234	260	p3	LGLSTLGELK	LGLSTLGELK	100
E234	260	p4	KASDVHEVR	KASNGHGELK	44,4
E235	318	p2	KVDWLTEK	TGLSTLGELK	14,3
E235	318	p3	LGLSTLGELK	LGLSTLGELK	100
E235	318	p5	KEEPSNNVK	FKEGHGELK	22,2
E245	293	p2	KVDWLTEK	KVDWLTEK	100
E245	293	p4	KASDVHEVR	KASDVMYVK	55,5
E245	293	p5	KEEPSNNVK	KEEPSPTEK	66,7
E345	278	p3	LGLSTLGELK	LGLSTLGELK	100
E345	278	p4	KASDVHEVR	KASNGHGELK	33,3
E345	278	p5	KEEPSNNVK	FKEGHGELK	22,2

O procedimento para os testes multiplex 3x3 foi detalhado no item 5.4 e representado na Figura 24. Os espectros multiplex simulados 3x3 apresentados na Tabela 29 foram usados para teste das heurísticas e posterior avaliação dos resultados. Os testes restringiram-se ao uso das três heurísticas que apresentaram melhores resultados nos espectros simulados 2x2, quais sejam, H3, H4 e H6.

A Tabela 30 mostra uma marcação com um X nas heurísticas que acertaram no sequenciamento *de novo* dos peptídeos menos intensos. As marcações em verde indicam as sequências do segundo peptídeo que já haviam sido sequenciadas corretamente na primeira tentativa antes de aplicar as heurísticas (vide Tabela 29).

Tabela 30: Peptídeo menos intenso sequenciado corretamente usando diferentes heurísticas multiplex. Estão marcados com fundo verde os peptídeos menos intensos sequenciados corretos originalmente. A marcação com um X representa que houve um sequenciamento *de novo* correto do espectro para a heurística

Espectro simulado	Precursor	Heurística		
		H3	H4	H6
E123	p1	X	X	X
E123	p2	X	X	
E123	p3			
E124	p1	X	X	X
E124	p2	X		
E124	p4			X
E125	p1	X	X	X
E125	p2			
E125	p5	X		
E134	p1	X	X	X
E134	p3	X	X	
E134	p4			X
E135	p1	X	X	X
E135	p3	X	X	
E135	p5			
E145	p1	X	X	X
E145	p4	X	X	X
E145	p5	X	X	
E234	p2	X	X	X
E234	p3		X	X
E234	p4	X		
E235	p2	X	X	X
E235	p3		X	X
E235	p5			
E245	p2	X	X	X
E245	p4	X	X	
E245	p5		X	
E345	p3	X	X	X
E345	p4	X	X	X
E345	p5			

As heurísticas H3, H4 e H6 mostraram ganhos no sequenciamento *de novo*, sendo que H4 mostrou-se ligeiramente melhor quando conseguiu sequenciar corretamente os três peptídeos com o espectro E145 e E245, enquanto que a heurística H3 conseguiu o mesmo feito somente para o espectro E145. Tanto H3 quanto H4 sequenciaram corretamente 20 espectros. Já H6 sequenciou corretamente 16 espectros.

As três estratégias de abatimento H3, H4 e H6 foram testadas em espectros multiplex reais.

7.3 Sequenciamento *de novo* usando espectros multiplex reais

Assim, as três melhores heurísticas multiplex foram usadas aqui, a saber, H3, H4 e H6, para avaliar a resposta sobre espectros multiplex reais. Os terceiro e quarto conjuntos de espectros foram usados nesta seção. O terceiro para identificar os peptídeos da amostra e o quarto para testar os programas para sequenciamento *de novo* nos espectros adquiridos em janela de fragmentação ampla (20 *m/z*) e descomplexados pela heurística proposta.

Para determinar os possíveis peptídeos presentes nos espectros multiplex reais, identificaram-se os peptídeos dos espectros da mesma amostra de tireóide na corrida no terceiro conjunto de espectros (DDA com janela de 2 *m/z*) usando o programa Comet. Todas as identificações do Comet com *score* acima de 1.0 e deltaCN acima de 0.1 foram consideradas. Isso possibilitou saber, *a priori*, quais sequências de aminoácidos poderiam ser encontradas nos espectros multiplex reais.

Em cada janela ampla centrada nos precursores de carga +2 que tenham sido identificados pelo Comet, procurou-se outros peptídeos alvo de carga +2 presentes nesta mesma janela de fragmentação e também identificados pelo Comet. Selecionou-se, então, os espectros multiplex e seus peptídeos alvo co-fragmentados na mesma janela e identificados pelo Comet. Os limites de procura dos peptídeos nos espectros MS^1 são +/- 10 *m/z* da massa dos precursores. Assim, exemplificando, para uma massa de precursor no espectro MS^2 de 652,13 Da, o limite inferior é 642,13 Da, enquanto que o limite superior é 662,13 Da. As massas de todos os peptídeos de carga +2 encontrados nesse intervalo são consideradas como peptídeos co-fragmentados no espectro multiplex MS^2. A Tabela 31 mostra a quantidade de janelas com diferentes números de peptídeos ionizados de carga +2 localizados nos intervalos, independentemente da existência ou não dos picos da série **y**.

Tabela 31: Número de janelas existentes considerando a quantidade de peptídeos de carga +2 existentes na janela calculada de 20 *m/z*

Quantidade de peptídeos na janela	Número de espectros
2	3930
3	2671
4	1309
5	525
6	158
7	51
8	7

Quanto maior a quantidade de peptídeos de carga +2 encontrados na janela, maior será a complexidade, pois os íons fragmentados de todos os peptídeos são registrados no mesmo espectro MS^2. Peptídeos alvo mais intensos teoricamente gerariam fragmentos mais intensos. Caso um peptídeo da janela tenha intensidade muito elevada enquanto outro tenha intensidade baixa, os íons fragmentos mais intensos podem predominar impedindo o sequenciamento *de novo* do peptídeo de menor intensidade, por isso a necessidade das heurísticas de descomplexação do espectro para cada peptídeo alvo.

Dada a complexidade destes espectros, os testes realizados se limitaram aos espectros MS^2 multiplex com 2 ou 3 peptídeos co-fragmentados e que continham todos os picos dos fragmentos tipo **y** da sequência esperada para a massa do peptídeo. Além disso, estes peptídeos alvo foram identificação pelo Comet na aquisição com janelas de isolamento de 2 *m/z*, de acordo com os critérios de qualidade estabelecidos. Esse procedimento faz-se necessário para garantir a existência dos caminhos nos espectros multiplex, o que limitou a quantidade de espectros multiplex válidos para os testes. Os espectros multiplex, assim anotados e selecionados, contendo dois peptídeos totalizaram 27 (54 peptídeos) e os espectros multiplex contendo 3 peptídeos totalizaram 2 (6 peptídeos).

A Tabela 32 mostra a quantidade de peptídeos sequenciados corretamente pelo DNbuilder com e sem uso das heurísticas H3, H4 e H6 nestes espectros. Os mesmos espectros foram sequenciados também pelos programas usados neste trabalho, Peaks, pNovo e Novor.

Tabela 32: Número de sequenciamentos *de novo* corretos de espectros multiplex com dois peptídeos para os diferentes programas usados nos testes, considerando ou não as heurísticas multiplex

Programa	Número de sequências corretas encontradas SEM heurística multiplex	Número de sequências corretas encontradas usando heurísticas multiplex		
		H3	H4	H6
27 espectros com 2 peptídeos				
DNbuilder	12	14	14	14
Peaks	16	18	18	18
pNovo	5	6	7	7
Novor	15	15	17	17

Na Tabela 32 podemos ver que as heurísticas produziram melhora discreta nos sequenciamentos *de novo* de forma bem diferenciada para cada programa para sequenciamento *de novo* usado sobre a mesma descomplexação. O desempenho do pNovo ficou bem distante dos outros algoritmos. O Peaks, comercial, ficou bem a frente de todos e o Novor recentemente lançado, teve desempenho melhor nas heurísticas H4 e H6, enquanto não obteve melhora no sequenciamento *de novo* com a heurística H3. O programa Peaks foi o que mais se destacou, conseguindo sequenciar corretamente a maior quantidade de peptídeos, considerando ou não as heurísticas.

A Tabela 33 mostra a quantidade de espectros multiplex com dois peptídeos em que todos os peptídeos foram sequenciados corretamente pelos diferentes programas, com ou sem heurística multiplex.

Tabela 33: Número de espectros multiplex em que todos os dois peptídeos foram sequenciados corretamente pelos diferentes programas testados, considerando ou não as três diferentes heurísticas multiplex

Programa	Número de espectros em que TODOS os peptídeos foram sequenciados corretamente sem heurística	Número de espectros em que TODOS os peptídeos foram sequenciados corretamente usando diferentes heurísticas multiplex		
		H3	H4	H6
27 espectros com 2 peptídeos				
DNbuilder	0	2	2	0
Peaks	1	3	3	1
PNovo	0	0	0	0
Novor	0	0	0	0

Podemos observar que apenas dois programas conseguiram sequenciar todos os peptídeos de um espectro multiplex, a saber Peaks e DNbuilder. Entre esses dois, sem a utilização de qualquer heurística o programa Peaks já realizava o sequenciamento *de novo* de maneira correta para 1 espectro multiplex, enquanto o programa DNbuilder não obteve nenhum êxito. Usando as diferentes heurísticas os ganhos foram equivalentes ambos com dois acertos, sendo que um dos espectros sequenciados corretamente foi comum aos dois programas. DNbuilder e Peaks perderam eficiência na heurística H6.

Cabe esclarecer que os resultados dos dois espectros contendo três peptídeos não foram apresentados nas Tabelas 28 e 29 porque não foi possível sequenciar corretamente nenhum dos seis possíveis peptídeos por nenhum dos programas utilizados. Tal fato comprova a complexidade de lidar com os fragmentos presentes nos espectros multiplex.

Ainda analisando os resultados na utilização dos espectros multiplex, contabilizou-se o posicionalmente de cada aminoácido correto do sequenciamento *de novo*. Isso permitiu saber o percentual geral de acerto em termos não de sequências corretas, mas sim de resíduos de aminoácidos corretamente sequenciados. A Tabela 34 mostra a quantidade de aminoácidos sequenciados corretamente pelos diferentes programas, considerando ou não as diferentes heurísticas, usando espectros multiplex com dois ou três peptídeos.

Tabela 34: Número de resíduos de aminoácidos sequenciados corretamente pelos diferentes programas, considerando ou não as diferentes heurísticas

Programa	Número deresíduos de aminoácidos sequenciados corretamente sem heurísticas	Número de resíduos de aminoácidos sequenciados corretamente usando diferentes heurísticas multiplex		
		H3	H4	H6
Espectros multiplex contendo 2 peptídeos (Somatório de 587 aminoácidos em todas as sequências identificadas pelo Comet)				
DNbuilder	216	247	253	272
Peaks	274	285	303	319
pNovo	207	195	211	215
Novor	205	201	213	230
Espectros multiplex contendo 3 peptídeos (Somatório de 56 aminoácidos em todas as sequências identificadas pelo Comet)				
DNbuilder	15	13	17	21
Peaks	23	21	27	27
pNovo	22	22	23	25
Novor	8	7	10	12

Na Tabela 34 podemos observar que todos os programas obtiveram algum sucesso na utilização das heurísticas, contabilizando maior quantidade de resíduos de aminoácidos na posição correta. A heuristica que mais se destacou foi a H6 para todos os programas, mostrando ser mais eficiente avaliando isoladamente os aminoácidos. Em termos percentuais, para os espectros contendo dois peptídeos, o programa Peaks conseguiu sequenciar corretamente 54,3% (319 de 587), enquanto o segundo mais eficiente foi o programa DNbuilder, 46,3% (272 de 587), seguido por Novor, 39,2% (230 de 587) e, por fim, pNovo, 36,6% (215 de 587).

Usar espectros com três peptídeos confirmou os resultados com dois peptídeos, onde mostra que o programa Peaks conseguiu o maior número de aminoácidos sequenciados na posição correta, enquanto o programa DNbuilder aparece como o segundo mais eficiente. A surpresa nos resultados usando espectros com três peptídeos é que o programa pNovo superou o Novor, sequenciando 25 aminoácidos na posição correta, enquanto o Novor sequenciou 12.

Os testes usando espectros multiplex reais adquiridos em janelas mais ampla, 20 *m/z*, não foram tão bons quanto usando espectros multiplex simulados. Para o programa DNbuilder observa-se um fato importante a se destacar decorrente da escolha da busca pela série **y**, que é a ausência constante do pico de massa nos espectros da lisina e da

arginina, que representa o último aminoácido da sequência tripsinada. Isso acontece porque os íons-fragmentos foram registrados a partir da *m/z* 200, possivelmente por escolha do operador do equipamento.

Outro fato importante para se destacar é a baixa qualidade dos espectros, observada pelo número elevado de identificações Comet abaixo de 1,5, mostrado na Tabela 5, 14766 espectros, de um total de 16595 espectros MS^2.

A identificação Comet fraca indica a pouca qualidade dos espectros multiplex reais, o que dificulta ou impede o sequenciamento *de novo*, como salientado na literatura [2]. A quantidade muito elevada de fragmentos existentes nos espectros multiplex do conjunto testado impediram também a determinação dos pares complementares nos espectros MS^2, na forma como foi sugerida recentemente por KRYUCHKOV [2].

8 CONCLUSÕES E PROPOSTAS FUTURAS

Esta tese apresentou uma abordagem para sequenciamento *de novo* de peptídeos em espectros MS^2 multiplex, adquiridos em espectrômetros de massa, que contêm fragmentos de mais de um peptídeo na mesma janela de fragmentação. Foi desenvolvido um algoritmo para sequenciamento *de novo* de peptídeos, denominado DNbuilder, que sequencia a série de aminoácidos considerando os ions-fragmentos do tipo **y**. O problema foi modelado através de grafos e adotou-se o algoritmo de busca DFS (*Depth-first search*) para se obter as sequências candidatas dos peptídeos. A pontuação, simples, considera somente a intensidade dos ions, picos do espectro de massa MS^2. O processo se inicia com a identificação da razão massa sobre carga (*m/z*) dos íons peptídeos monoisotópicos de carga +2 presentes na janela selecionada para fragmentação do primeiro espectro de massa, MS^1. A metodologia multiplex fundamenta-se na alteração das intensidades dos picos fragmentos do segundo espectro, MS^2, para cada novo peptídeo alvo a ser sequenciado da janela, atenuando picos do espectro correspondentes a fragmentos do tipo **y**/**b** de peptídeos já identificados. Os programas de sequenciamento *de novo* usados para validar a metodologia de sequenciamento *de novo* foram o Peaks 8, pNovoPlus e Novor 1.3.489, e assim comparados com DNBuilder. Espectros de uma amostra de tireoide adquiridos em janelas de 20 *m/z* foram usados nos testes de avaliação da metodologia multiplex. Os resultados mostram que a metodologia, mesmo que simples, melhora o sequenciamento *de novo* dos peptídeos presentes nos espectros multiplex MS^2, aumentando o número de acertos de aminoácidos corretos das sequências, mostrando que há um caminho possível para o sequenciamento *de novo* de peptídeos em espectros multiplex para janelas amplas.

A adoção do sequenciamento *de novo* através da série de fragmentos **y**, com a utilização de arestas duplas penalizadas e com pontuação baseada apenas nas intensidades dos picos, com pós-processamento de pontuação, teve bons resultados. A comprovação dessas escolhas foi o fato do programa DNbuilder conseguir sequenciar *de novo* muitos espectros que outros programas não conseguiram.

O sequenciamento *de novo* de peptídeo é profundamente dependente da qualidade dos espectros usados. Nenhum bom resultado será obtido a partir de espectros de má qualidade. Três conjuntos de espectros foram usados para validar a metodologia

de sequenciamento *de novo*. O terceiro conjunto de espectros usado nos testes ratifica a necessidade da qualidade dos espectros, pois foi a que obteve o maior número de espectros com identificação Comet com *score* abaixo de 1,5 e a que apresentou a menor taxa de acerto, mostrada na Tabela 18. Uma melhor qualidade dos espectros permite, ainda, que se use uma tolerância do erro dos fragmentos muito menor, 0,02 Da, que faz com que se diminua a complexidade do problema. Essa redução da complexidade faz com que o processamento seja mais rápido e diminui as chances do algoritmo de busca no grafo errar o caminho.

A implementação de um programa, DNbuilder, resultou em conclusões interessantes, como a implicação do uso de arestas duplas no grafo aumentou a quantidade de acertos. Outro fato interessante observado foi a de que o grafo construído a partir da série **y** parece contribuir no sequenciamento *de novo* de um volume considerável de espectros que os algoritmos Peaks, pNovo e Novor não atingem. Possivelmente devido ao fato da maior presença dos picos da série **y** no espectro, aliado ao fato dessa série ser mais abundante, isto é, picos mais intensos. Desenvolver um programa próprio também possibilitou investigar a influência das evidências na pontuação, onde os testes mostraram que a utilização usando apenas as intensidades dos picos da série **y** é bastante eficiente.

A dificuldade no sequenciamento *de novo* de espectros multiplex foi a obtenção de espectros comprovadamente multiplex, o que resultou na criação de espectros multiplex simulados. Os testes em espectros multiplex simulados apresentaram resultados animadores para as heurísticas multiplex, o que não se repetiu usando espectros multiplex reais, mesmo que os resultados mostrassem alguma melhora nas sequências.

A distância dos resultados usando espectros MS^2 simulados e reais pode ter ocorrido por causa das diferentes formas de criação dos mesmos. Os simulados resultaram na união de todos os picos entre dois ou três espectros. Já os espectros MS^2 reais registraram todos os fragmentos dos peptídeos eluídos na mesma janela de aquisição, que aumenta consideravelmente a quantidade de fragmentos do espectro. E complicando ainda mais o experimento multiplex desse trabalho, a janela ampla de aquisição possuía um tamanho de 20 *m/z*, o que aumentou potencialmente a quantidade dos fragmentos registrados nos espectros MS^2. Ainda assim, mesmo com os espectro MS^2 com picos em demasia, a heurística multiplex conseguiu melhorar o

sequenciamento *de novo*, apesar de estar longe do que se espera. Não há na literatura estudos de sequenciamento *de novo* usando espectros multiplex com janelas de aquisição tão ampla, com espectros tão complexos.

Futuramente, uma avaliação das frequências sobre a presença de íons tipo **b** ou **y** nos espectros pode indicar qual série seria prioritária na criação do grafo a partir de picos de espectro MS^2, analisando inclusive, possível correlação com o tipo de fragmentação CID/HCD, a exemplo da tentativa de alguns autores em usar estatísticas obtidas em espectros de treinamento que ajude na tomada de decisão sobre íon a ser sequenciado, pontuação, etc.

Para o sequenciamento *de novo* de peptídeos, uma outra ideia para continuidade de trabalho é a realização de testes de uma construção de grafo mista, usando ambas as séries de picos **b** e **y**. Outra opção viável seria tentar sequenciar usando a série **b** apenas quando não houver resultado satisfatório para a série **y**. De qualquer forma, considerar a aresta dupla será sempre uma boa opção, visto que melhorou consideravelmente o resultado do sequenciamento *de novo* apresentado aqui, mostrando confiabilidade. Talvez haja a necessidade de sofisticar o sistema de pontuação, como usar Picos Complementares-CP [2] para melhorar a busca por causa da menor intensidade dos picos da série **b**, e não se trata de tarefa simplificada. Sabe-se que a maioria dos algoritmos leva em conta treinamentos de um volume considerável de espectros anotados para melhorar a interpretação e desempenho dos algoritmos, mas a tendência de se adquirir cada vez melhores espectros deve derrubar esta necessidade, o que ajudaria na sofisticação do sistema de pontuação.

Independentemente das melhorias futuras do programa DNbuilder, pretende-se realizar sequenciamento *de novo* de peptídeos usando espectros multiplex com janelas pequenas para viabilizar novas heurísticas mais eficientes do que as testadas neste trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] FACCIN, M.; BRUSCOLINI, P.; "MS/MS Spectra Interpretation as a Statistical–Mechanics Problem". **Analytical Chemistry** *85* (10), 4884-4892, 2013.

[2] KRYUCHKOV, F.; VERANO-BRAGA, T.HANSEN, T.A.; SPRENGER, R.R.; KJELDSEN, F.; "Deconvolution of Mixture Spectra and Increased Throughput of Peptide Identification by Utilization of Intensified Complementary Ions Formed in Tandem Mass Spectrometry". **Journal of Proteome Research**, 12, 3362-3371, 2013.

[3] FENN, J.B.; MANN, M.; MENG, C.K.; WONG, S.F.; WHITEHOUSE, C.M.; "Electrospray ionization for mass spectrometry of large biomolecules". **Science**, vol. 246, no. 4926, pp. 64–71, 1989.

[4] STEEN, H.; MANN, M.; "The ABC's (and XYZ's) of peptide sequence". **Molecular & Cellular Biology**, 5, 699-711, 2004.

[5] PERRY, R.H.; COOKS, R.G.; Noll, R.J.; "Orbitrap mass spectrometry: instrumentation, ion motion and applications". **Mass Spectrometry Reviews**. 27:661–99, 2008.

[6] GOOD, M.G.; MARIN-VICENTE, C.; ZUBAREV, D.M.; "Are the majority of a_2-ions cyclic?", **Physical Chemistry Chemical Physics**, 12, 13372-13374, 2010.

[7] MATTAUCH, J.; HERZOG, R.; "Mass spectrography". **Z. Physik**, 89:786, 1934.

[8] JOHNSON, E.G.; NIER, A.O.; "Angular aberrations in sector shaped electromagnetic lenses for focusing beams of charged particles". **Physical Review**, 91:10, 1953.

[9] BRENTON, A.G.; GODFREY, A.R.; "Accurate mass measurement: terminology and treatment of data". **Journal of the American Society for Mass Spectrometry**, 21:1821-1835, 2010.

[10] LANÇAS, F.M.; "A cromatografia líquida moderna e a espectrometria de massas: Finalmente 'compatíveis'?II. A escolha do analisador de massas". **Scientia Chromatographica**, 5(1):27-46, 2013.

[11] FRANK, A.M.; SAVITSKI, M.M.; NIELSEN, M.L.; ZUBAREV, R.A.; PEVZNER, P.A.; "*De novo* peptide sequencing and identification with precision mass spectrometry". **Journal of Proteome Research**, 6:114–123, 2007.

[12] HASSELL, K.M.; LeBLANC, Y.; McLUCKEY, S.A.; "Chemical Noise Reduction via Mass Spectrometry and Ion/Ion Charge Inversion: Amino Acids". **Analytical Chemistry**, 83, 3252–3255, 2011.

[13] SOUZA, L. M.; *Aplicações da espectrometria de massas e da cromatografia líquida na caracterização estrutural de biomoléculas de baixa massa molecular*. Tese de Mestrado. Programa de Pós-Graduação em Bioquímica da UFPR, Curitiba, 2008.

[14] K.F. CHONG, K.F.; LEONG, H.W.; "Tutorial on *de novo* peptide sequencing using MS/MS mass spectrometry". **Journal of Bioinformatic and Computational Biology**, 10, p. 1231002, 2012.

[15] ROEPSTORFF, P.; FOHLMAN, J.; "Proposal for a common nomenclature for sequence ions in mass spectra of peptides". **Biomedical Mass Spectrometry** 11: 601, 1984.

[16] SANTOS, L.D.; "Estratégias e aplicações da Espectrometria de Massas". IN: Workshop Avanços na Engenharia de Proteínas e Peptídeos. Fundação Oswando Cruz, 13:17, 2008.

[17] CANTÚ, M.D.; CARRILHO, E.;WULFF, N.A.; PALMA, M.S.; "Sequenciamento de Peptídeos Usando Espectrometria de Massa: um guia prático". **Química Nova**, Vol.31, No. 3, 669-675, 2008.

[18] SAKURAI, T.; MATSUO, T.; MATSUDA, H.; KATAKUSE, I.; "PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data". **Biomedical Mass Spectrometry**, 11(8), 396-399, 1984.

[19] HAMM, C. W.; WILSON, W. E.; HARVAN, D. J.; "Peptide sequencing program". **Computer Applications in the Bioscience**, 2, 115-118, 1986.

[20] ISHIKAWA, K.; NIVA, Y.; "Computer-Aided Peptide Sequencing by Fast Atom Bombardment Mass Spectrometry". **Biomedical and Environmental Mass Spectrometry**,13, 373-380, 1986.

[21] SIEGEL, M. M.; BAUMAN, N.; "An Efficient Algorithm for Sequencing Peptides Using Fast Atom Bombardment Mass Spectral Data". **Biomedical and Environmental Mass Spectrometry**, 15, 333-343, 1988.

[22] BARTELS, C.; "Fast Algorithm for Peptide Sequencing by Mass Spectroscopy". **Biomedical and Environmental Mass Spectrometry**, 19, 363-368, 1990.

[23] TAYLOR, J.A.; JOHNSON, R.S.; "Sequence Database Searches via *De Novo* Peptide Sequencing by Tandem Mass Spectrometry". **Rapid Communication in Mass Spectrometry**, 11, 1067-1075, 1997.

[24] DANCÍK, V., ADDONA, T.A., CLAUSER, K.R., VATH, J.E., PEVZNER, P.A.; "*De novo* peptide sequencing via tandem mass spectrometry". **Journal of Computational Biology**, 6 (3/4):327-342, 1999.

[25] MA, B., ZHANG, K., HENDRIE, C., LIANG, C., LI, M., DOHERTY-KIRBY, A., LAJOIE, G.; "PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry". **Rapid Communication in Mass Spectrometry**, vol. 17, pp. 2337-42, 2003. (http://www.bioinfor.com/peaks/tutorials/denovo.html)

[26] CHEN, T.; KAO, M-Y.; TEPEL, M.; RUSH, J.; CHURCH, G.M.; "A Dynamic Programming Approach to *De Novo* Peptide Sequencing via Tandem Mass Spectrometry". **Journal Computational Biology**. 2001, 8(3), 325-337.

[27] FRANK, A., PEVZNER, P.A.; "PepNovo: *de novo* peptide sequencing via probabilistic network modeling". **Analytical Chemistry** 77, 964-73, 2005.

[28] HINES, W.M., FALICK, A.M., BURLINGAME, A.L., AND GIBSON, B.W.; "Pattern-Based Algorithm for Peptide Sequencing from Tandem High Energy Collision-Induced Dissociation Mass Spectra". **Journal of the American Society for Mass Spectrometry**, 3, 326-336, 1992.

[29] MARCH, R.E.; Introduction to Quadrupole Ion Trap Mass Spectrometry. **Journal of Mass Spectrometry**, 32:351-369, 1997.

[30] BAGINSKY, S., CIELIEBAK, M., GRUISSEM, W.,KLEFFMANN, T., LIPTAK, Z., MULLER, M., PENNA, P.; "AuDeNS - A Tool for Automatic *De Novo* Peptide Sequencing". Technical Report no. 383, **ETHZ**, 2002.

[31] MO, L., DUTTA, D., WAN, Y., CHEN T.; "MSNovo: a dynamics programming algorithm for *de novo* peptide sequencing via tandem mass spectrometry". **Analytical Chemistry**, 79(13):4870-8, 2007.

[32] TABB, D.L., MA, Z., MARTIN, D.B., HAM, A.L, CHAMBERS, M.C.; "DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring". **Journal of Proteome Research**, 7(9):3838-46, 2008.

[33] CHI, H., SUN, R., YANG, B., SONG, C., WANG, L., LIU, C., FU, Y., YUAN, Z., WANG, H., HE, S., DONG, M.; "pNovo: *de novo* peptide sequencing and identification using HCD spectra". **Journal of Proteome Research** 9, 2713-24, 2010.

[34] LI, H., LIU, C., RWEBANGIRA, M., BURGE, L., SOULTHERLAND, W.; "Rapid Generation of Peptide Sequence Tags with a Graph Search Algorithm". **Bioinformatics and Biomedicine Workshop**, p. 251-254, 2011.

[35] JEONG, K., KIM, S. & PEVZNER, P.A.;"UniNovo: a universal tool for *de novo* peptide sequencing". **Bioinformatics** 29, 1953-1962, 2013.

[36] DIESTEL, R.; *Graph theory*. Vol. 173. 4 ed. Heidelberg, Springer-Hidelberg, 2010.

[37] ZHANG, Y.; FONSLOW, B.R.; SHAN, B.; BEAK, M.; YATES III, J.R.; "Protein analysis by shotgun/bottom-up proteomics". **Chemical Reviews**, 113(4): 2343-2394, 2013.

[38] DEUTSCH, E. W.; "File formats commonly used in mass spectrometry proteomics". **Molecular & Cellular Proteomics** 11, 1612-1621, 2012.

[39] KELLER, A.; PURVINE, S.; NESVIZHSKII, A. I.; STOLYAR, S.; GOODLETT, D.R.; KOLKER, E.; "Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis". **OMICS** 2002, 6(2), 207-212, 2002.

[40] MA, B.; JOHNSON, R.; "*De novo* sequencing and homology searching". **Molecular & Cellular Proteomics**, O111.014902, 2011.

[41] PEVTSOV, S.; FEDULOVA, I.; MIRZAEI, H.; BUCK, C.; ZHANG, X.; "Performance evaluation of existing *de novo* sequencing algorithms". **Proteome Research**, 5(11), 3018–28, 2006.

[42] LEPREVOST, F.V. *et* al; "PepExplorer: a similarity-driven tool for analysing *de novo* sequencing results". **Molecular & Cellular Proteomics**, M113.037002, 2014.

[43] TARJAN, R.E.; "Depth-first-search and linear graph algorithm". SIAM **Journal on Computing**, 2:146–160, 1972.

[44] LANE, L.; ARGOUD-PUY, G.; BRITAN, A.; CUSIN, I.; DUEK, P.D.; EVALET, O.; GATEAU, A.; GAUDET, P.; GLEIZES, A.; MASSELOT, A. *et* al.; "neXtProt: a knowledge platform for human proteins". **Nucleic Acids Research**, 40, D76–D83, 2012.

[45] GUTHALS, A.; CLAUSER, K.R.; FRANK, A.M.; BANDEIRA, N.; "Sequencing-Grade *De Novo* Analysis of MS/MS Triplets (CID/HCD/ETD) From Overlapping Peptides". **Journal of Proteome Research**, 12, 2846-2857, 2013.

[46] MEDZIHRADSZKY, K.F.; CHALKLEY, R.J. "Lessons in *De Novo* Peptide Sequencing by Tandem Mass Spectrometry". **Wiley**, thrid edition, 2013.

[47] MA, B.; "Challenges in computational analysis of mass spectrometry data for proteomics". **Journal of Computer Science and Technology**, 25(1): 1 Jan. 2010.

[48] ENG, J.K.; JAHAN, T.A.; HOOPMANN, M.R.; Comet: "An open-source MS/MS sequence database search tool". **Proteomics**, 13, 22-24, 2013.

[49] MA, B.; "Novor: Real-Time Peptide *De Novo* Sequencing Software". **Journal of The American Society for Mass Spectrometry**, vol.26, issue 11, 1885-1894, 2015.

[50] GORSHKOV, V.; VERANO-BRAGA, T.; KJELDSEN, F.; "SuperQuant: A Data Processing Approach to Increase Quantitative Proteome Coverage". **Analytical Chemistry**, 87, 6319-6327, 2015.

[51] GORSHKOV, V.; HOTTA, S.Y.K.; VERANO-BRAGA, T.; KJELDSEN, F.; "Peptide *de novo* sequencing of mixture tandem mass spectra". **Journal of Proteomics**, 16, 2470-2479, 2016.

[52] COOKS, R. G.; Rockwood, A. L. "The 'Thomson'. A suggested unit for mass spectroscopists". ***Rapid Communications in Mass Spectrometry***. **5** (2): 93, 1991.

[53] McHUGH, L.; ARTHUR, J.W. “Computational Methods for Protein Identification from Mass Spectrometry Data”. **PLoS Computational Biology**, 4 (2), 2008.

[54] PRINCE, J.T.; CARLSON, M.W.; WANG, R.; MARCOTTE, E.M. “The Need for a Public Proteomics Repository”. **Nature Biotechnology**, 22, 4, 2004.

[55] CARVALHO, P.C.; FISHER, J.S.G.; CHEN, E.I.; YATES, J.R.; BARBOSA, V.C. “PatternLab for proteomics: a tool for differential shotgun proteomics”. **BMC Bioinformatics**, 9, 316, 2008.

[56] VESSECCHI, R; LOPES, N.P.; GOZZO, F.; DORR, F.A.; MURGU, M.; LEBRE, D.T.; ABREU, R.; BUSTILLOS, O.V.; RIVEROS, J.M. “Nomenclaturas de Espectrometria de Massas em Língua Portuguesa”. **Química Nova**, vol.34, 10, 1875-1887, 2011.

[57] CHONG, K.F.; LEONG, H.W.”Tutorial on *De Novo* Peptide Sequencing Using MS/MS Mass Spectrometry”. **Journal of Bioinformatics and Computational Biology**, vol.10, 6, 1231002, 2012.

[58] CANTÚ, M.D.; CARRILHO, E.; WULLF, N.A.; PALMA, M.S.”Sequenciamento de Peptídeos Usando Espectrometria de Massas: Um Guia Prático”. **Química Nova**, vol.31, 3, 669-675, 2008.

[59] FALIC, A.M; HINES, W.M.; MEDZIHRADSZKY, K.F.; BALDWIN, M.A.; GIBSON, B.W. “Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass spectrometry”. **Journal of the American Society for Mass Spectrometry**, vol. 4, 882-893, 1993.

[60] NEDA, I; VLAZAN, P.; POP, R.O.; SFIRLOAGA, P.; GROZESCU,I.; SEGNEANU, A.E.; “Peptide and Amino Acids Separation and Identification from Natural Products”. In **Analitical Chemistry**, Chapter 6. Ed. Ira S. Krull, ISBN 978-953-51-0837-5, 2012.

[61] OLSEN, J.V.; SCHWARTZ, J.C.; GRIEP-RAMING, J. *et* al. “A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. **Molecular & cellular proteomics**: MCP, 8, 12, p.2759-2769, 2009.

[62] AQUINO, P.F. “Avaliação Proteômica da Margem de Ressecção de Pacientes com Câncer Gástrico por Espectrometria de Massas”. Tese. UFRJ-Instituto de Química, 2015.