UNIVERSIDADE FEDERAL DO PARANÁ

MALTON WILLIAM MACHADO CUNICO

JMSA: JAVA MASS SPECTROMETRY ANALYZER: FERRAMENTA PARA GERENCIAMENTO E ANÁLISE DE BANCO DE ESPECTROS DE MASSA PARA IDENTIFICAÇÃO DE MICRORGANISMOS

CURITIBA

2017

MALTON WILLIAM MACHADO CUNICO

JMSA: JAVA MASS SPECTROMETRY ANALYZER: FERRAMENTA PARA GERENCIAMENTO E ANÁLISE DE BANCO DE ESPECTROS DE MASSA PARA IDENTIFICAÇÃO DE MICRORGANISMOS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração Bioinformática.

Orientador: Prof. Dr. Leonardo Magalhães Cruz
Coorientador: Prof. Dr. Luciano Fernandes Huergo

CURITIBA

2017

Cunico, Malton William Machado
C972 JMSA: Java Mass Spectrometry Analyzer: ferramenta para gerenciamento e análise de banco de espectros de massa para identificação de microrganismos / Malton William Machado Cunico. – Curitiba, 2017.
52 f. : il.

Orientador: Prof. Dr. Leonardo Magalhães Cruz
Coorientador: Prof. Dr. Luciano Fernandes Huergo
Dissertação (mestrado) - Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Programa de Pós-Graduação em Bioinformática.
Inclui bibliografia

1. Bioinformática. 2. Microorganismos - Identificação. 3. Espectrometria de massa. 4. Java (Linguagem de programação de computador). I. Cruz, Leonardo Magalhães. II. Huergo, Luciano Fernandes. III. Universidade Federal do Paraná. IV. Título.

CDD – 576

TERMO DE APROVAÇÃO

MALTON WILLIAM MACHADO CUNICO

"JMSA: Java mass spectrometry analyzer: ferramenta para gerenciamento e análise de banco de espectros de massa para identificação de microrganismos"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Dr. Leonardo Magalhães Cruz
Universidade Federal do Paraná - UFPR

Dr. Dieval Guizelini
Universidade Federal do Paraná - UFPR

Dr. Rodrigo Luis Alves Cardoso
Bolsista PNPD/CAPES/Projeto Biologia Computacional - UFPR

Curitiba, 10 de março de 2017

AGRADECIMENTOS

Agradeço primeiramente a Deus segundo a minha crença.

Agradeço também a todos os meus familiares que apoiaram minhas decisões e me ajudaram nos momentos de aflição.

Agradeço pela orientação dos professores envolvidos, especialmente sou grato a paciência e compreensão do meu orientador, Prof. Dr. Leonardo Magalhães Cruz.

Sou grato ao Programa de Pós-Graduação em Bioinformática da UFPR e aos órgãos de fomento que junto ao colegiado do curso, colegas, professores e demais funcionários permitiram a realização deste.

RESUMO

A utilização da espectrometria de massa MALDI-TOF permite a identificação de microrganismos através da geração de espectros de massa, representando um perfil característico de sinais obtidos a partir de peptídeos ionizados de células inteiras ou extratos celulares. A comparação de espectros de massa obtidos de microrganismos desconhecidos contra um banco de dados de espectros de massa para microrganismos conhecidos, permite sua identificação. Essa utilização permite o usufruto de algumas vantagens frente a algumas limitações da técnica padrão, tal como agilidade na análise e redução de custos. Entretanto, faltam alternativas gratuitas aos softwares proprietários capazes de suprir características chaves na análise dos dados extraídos por tal técnica. Para auxiliar nessa análise nós criamos o JMSA, que é uma ferramenta de análise dos picos de um espectro de massa. O JMSA é capaz de facilitar a visualização comparativa, incluir dados descritivos de amostras. O JMSA também pode executar uma comparação de similaridade entre espectros selecionados. Desta forma o programa foi capaz de identificar um espectro, dado como desconhecido, em nível de espécie. O programa é capaz de ser executado consumindo poucos recursos nos principais sistemas operacionais que suportem Java, tal como MacOSX, Linux e Windows.

Palavras-chave: Espectrometria de massa, Identificação de microrganismos, Análise de espectros de massa, MALDI-TOF, Desenvolvimento de software, Java.

ABSTRACT

The use of MALDI-TOF mass spectrometry allows microorganism identification by generating mass spectra representing a characteristic profile of signals from ionized whole cell peptides or cell extracts. Comparing of mass spectra obtained from unknown microorganisms with a database of mass spectra for known microorganisms allows their identification. This usage allows the exploitation of some advantages over some limitations of the standard technique, such as agility in the analysis and reduction of costs. However, there is a lack of free alternatives to proprietary software capable of supplying key characteristics in its extracted data analysis. To assist in this analysis we have created the JMSA, which is a tool for analyzing the peaks of a mass spectrum. The JMSA is able to facilitate comparative visualization as well as include descriptive sample data. JMSA can also perform a similarity comparison of selected spectra. In this way the program was able to identify a spectrum, given as unknown, at species level. The program is able to run by consuming few resources on major operating systems that support Java, such as MacOSX, Linux and Windows.

Key-words: Mass spectrometry, Identification of microorganisms, Mass spectrum analysis software, MALDI-TOF, Software development, Java.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em pleno Século XXI, softwares usados para a bioinformática em geral, têm sido desenvolvidos por pesquisadores em universidades, pois investir no desenvolvimento de ferramentas que consigam responder melhor às necessidades específicas das pesquisas é um fator fundamental para o desenvolvimento científico e tecnológico de um país ("Com Ciência - Bioinformática", 2016).

Um exemplo disso é aliar a bioinformática à aplicação da espectrometria de massas, na caracterização de microrganismos. Esta técnica permite que a coleta de dados para identificação de microrganismos seja feita de forma muito mais simples, barata e rápida, do que os métodos bioquímicos convencionais automatizados, uma vez que o resultado não depende da metabolização de substâncias, além de que também produz uma quantidade de resíduos muito menor do que a produzida pelos métodos tradicionais. Entretanto os dados gerados e as análises são complexas, dependendo do desenvolvimento de aplicativos e algoritmos específicos para tornar a identificação confiável e rápida. Alguns sistemas de hardware e software integrados estão comercialmente disponíveis (inclusive no Brasil), incluindo as bases de dados: o Biotyper Bruker (Bruker Daltonics) e o Vitek® MS (bioMérieux).

Um dos maiores ganhos clínicos com a utilização da identificação por espectrometria de massa está na redução de horas gastas com o isolamento do organismo e a identificação bioquímica, calculado em até 48 horas. Sabemos que em casos de infecções sistêmicas, este tempo é extremamente precioso para o paciente. Muitos estudos têm demonstrado essa vantagem da identificação de fungos e de micobactérias através do MALDI-TOF.

Contudo, os bancos de dados e programas disponíveis no mercado são caros, fazendo com que haja uma demanda nessa área para a criação de programas que permitam manipulação e identificação de microrganismos por análise de espectros de massa.

Dessa forma, justifica-se nessa pesquisa criar um software para tal fim, uma vez que o mesmo facilitará o manuseio e a análise dos dados de espectrometria de massa por unir funcionalidades presentes individualmente nas diversas aplicações. Dentre essas funcionalidades e características podemos destacar:

• Armazenamento descritivo de espectros de massa

• Rápida reanálise de dados
• Comparação em pares
• Interface de usuário amigável
• Aplicação multiplataforma

1.1 JUSTIFICATIVA

A utilização da espectrometria de massa MALDI-TOF permite a identificação de microrganismos se comparados com espectros de microrganismos já previamente identificados. Essa utilização permite o usufruto de algumas vantagens frente a algumas limitações da técnica padrão.

O trabalho desenvolvido vai facilitar o manuseio e a análise dos dados de espectrometria de massa pois vai unir funcionalidades presentes individualmente nas diversas aplicações presentes. Dentre essas funcionalidades e características podemos destacar:

- Armazenamento descritivo de espectros de massa
- Rápida reanálise de dados
- Comparação em pares de espectros de massa
- Interface de usuário amigável
- Aplicação multiplataforma

Apesar dos benefícios da espectrometria de massa na coleta de dados para análise, tal análise voltada para identificação de microrganismos apresenta desvantagens, como: complexidade devido aos dados brutos, *hardware* caro, disponibilidade somente de software proprietário e de alto custo ou a falta de software livre que não suprem as necessidades ou características desejadas, justificando, desta forma, o desenvolvimento de software livre e algoritmos voltados para a comparação de espectros de massa e identificação de microrganismos.

1.2 OBJETIVOS

1.2.1 Objetivo geral

Desenvolver um aplicativo para identificação de microrganismos por análise de espectrometria de massa.

1.2.2 Objetivos específicos

Os objetivos específicos foram definidos como os seguintes:

- Desenvolver método para quantificar a semelhança entre espectros.
- Desenvolver uma interface gráfica intuitiva e amigável para o programa;
- Padronizar um arquivo de texto estruturado para gravação de dados de espectrometria de massa e informações de amostras biológicas.
- Disponibilizar visualização da análise feita.
- Executar um estudo de caso para testar a identificação de um espectro de massa.
- Avaliar a performance do programa

2 REVISÃO LITERÁRIA

2.1 ESPECTROMETRIA DE MASSA

A espectrometria de massa é uma técnica analítica para identificação dos componentes químicos presentes numa amostra.

Essa técnica data desde o fim do século 19, quando em 1886 Eugen Goldstein fez sua descoberta sobre os raios canais. Isso permitiu que em 1897 Joseph John Thomson fizesse a identificação de um corpúsculo de carga negativa muito menor do que o átomo, posteriormente sendo reconhecido como elétron e garantindo a Thomson o prêmio nobel em 1906. A descoberta de Goldstein também permitiu em 1899 a identificação por Wilhelm Wien de uma partícula de igual massa que o átomo de hidrogênio, posteriormente reconhecido como próton. Em 1899, Wien criou um dispositivo para separação de íons positivos de acordo com a sua relação carga massa. Já em 1912 Joseph John Thomson e Francis William Aston fizeram a descoberta dos isótopos de neon-20 e neon-22, sendo tal separação pioneira na técnica de espectrometria de massa. Em 1918 Arthur Jeffrey Dempster construiu o primeiro espectrômetro de massa moderno e em 1919 Aston construiu um espectrômetro de massa com a qual conseguiu identificar 212 isótopos naturais, descoberta que lhe rendeu o prêmio nobel em 1922. Algumas décadas depois Franz Hillenkamp e Michael Karas desenvolveram a técnica de “dessorção/ionização a laser assistida por matriz” (do inglês, *matrix-assisted laser desorption/ionization*), mais popularmente conhecida como MALDI. Essa técnica é até hoje uma das mais utilizadas técnicas para identificação de biomoléculas. Dois anos depois, em 1987, John Bennett Fenn desenvolveu a técnica de ionização por eletrospray (ESI), enquanto Koichi Tanaka desenvolveu a técnica de dessorção suave a laser (SLD, do inglês, *Soft laser desorption*). Ambos dividiram metade do prêmio nobel em 2002 por seus respectivos trabalhos. As principais datas e eventos para o desenvolvimento da espectrometria de massa são resumidos cronologicamente na Tabela 1.

Um espectrômetro de massa moderno consiste em três módulos (Croxatto; Prod'hom; Greub, 2012): a fonte ionizante, o analisador de massas e o detector, ou armadilha de íons. Já o seu funcionamento segue os seguintes passos: i) vaporização de uma amostra no espectrômetro de massa; ii) ionização da amostra e formação de partículas carregadas positivamente; iii) aceleração dessas partículas

em campo magnético; iv) computação da relação massa-carga; v) detecção de íons.

Tabela 1: Datas, eventos e pesquisadores importantes no desenvolvimento da espectrometria de massa.

Data	**Pesquisador**	**Evento**
1886	Eugen Goldstein	Raios canais—Descarga de íons positivos—Tubos de gás ionizado (Grayson, 2002)
1897	Joseph John Thomson	Identificação de corpúsculo negativo muito menor que o átomo (Thomson, 1897)
1898	Wilhelm Wien	Identificação de partícula de igual massa que o átomo de hidrogênio (du Bois, 1898)
1906	Joseph John Thomson	Nobel pelo descobrimento do elétron ("The Nobel Prize in Physics 1906", 2017)
1912	Joseph John Thomson e Francis William Aston	Descoberta dos isótopos de neon-20 e neon-22 (Thomson, 1912) (Falconer, 1997)
1918	Arthur Jeffrey Dempster	Construção do primeiro espectrômetro de massa moderno (Dempster, 1918)
1919	Francis William Aston	Desenvolvimento de um espectrômetro de massa e de 212 isótopos naturais (Aston, 1919)
1922	Francis William Aston	Nobel pela descoberta de grandes quantidades de isótopos em elementos não radioativos ("The Nobel Prize in Chemistry 1922", 2017)
1984	Masamichi Yamashita and John Bennett Fenn	Desenvolvimento da técnica de ionização por eletrospray (ESI) (Yamashita; Fenn, 1984)
1985	Franz Hillenkamp e Michael Karas	Desenvolvimento da técnica de ionização MALDI (Karas; Bachmann; Hillenkamp, 1985)
1988	Koichi Tanaka	Desenvolvimento da técnica de dessorção suave a laser (SLD, *Soft laser desorption*) (Tanaka et al., 1988)
2002	John Bennett Fenn e Koichi Tanaka	Nobel pelos trabalhos com ESI e SLD, respectivamente ("The Nobel Prize in Chemistry 2002", 2017)

FONTE: O autor (2017)

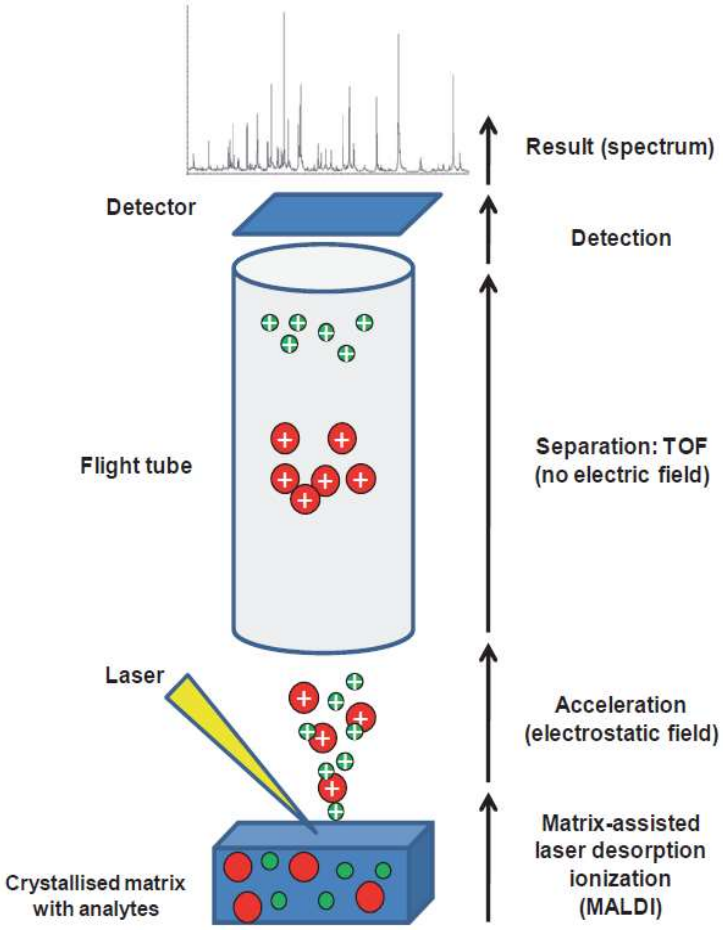
2.2 ESPECTROMETRIA DE MASSA POR MALDI-TOF

Este tipo de espectrometria utiliza a dessorção/ionização a laser assistida por

matriz (MALDI) como forma de ionização e o tempo de voo (ToF, do inglês, *time of flight*) como separador de massas e cargas.

A técnica MALDI, é uma técnica de ionização que permite a análise de amostras biológicas líquidas ou sólidas através da mistura com composto protetor/ionizador chamado de matriz. Essa matriz permite que sob indução de laser UV, uma amostra não ionizável seja ionizada e, por fim, possibilite sua análise. A matriz tem também uma função protetora da amostra, impedindo que seja degradada, mas permitindo sua ionização (Figura 1).

Figura 1: Base de funcionamento de espectrômetros de massa MALDI-ToF



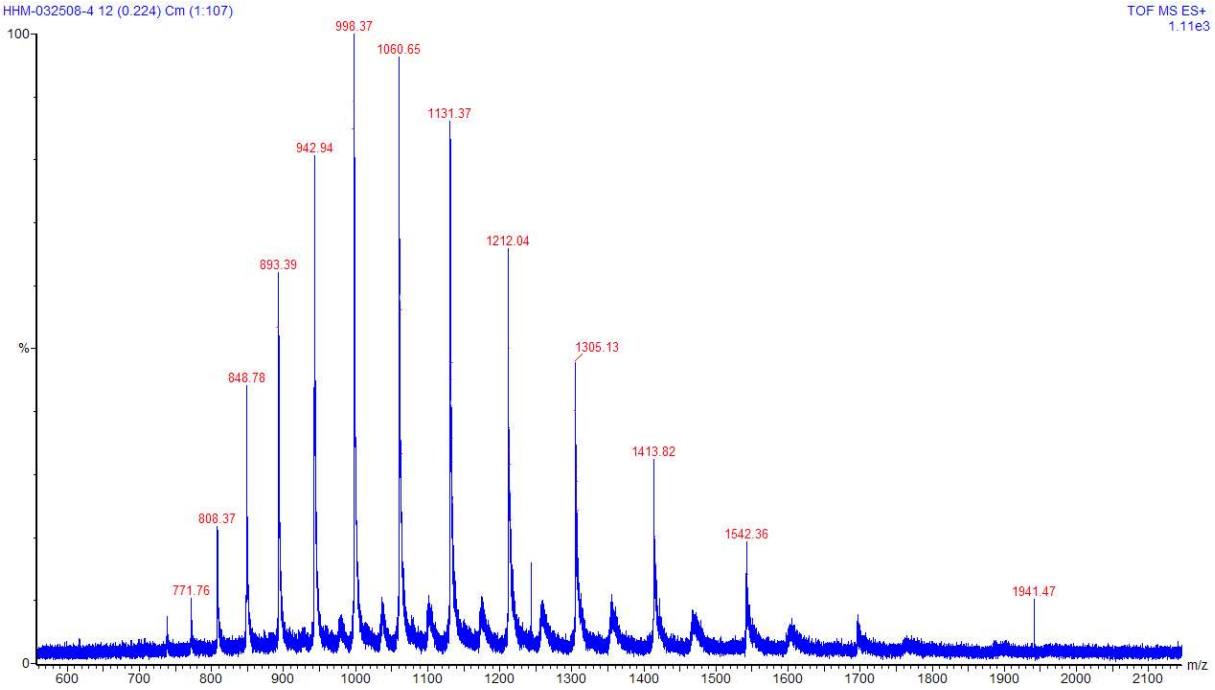
FONTE: Croxatto; Prod'homGreub, 2012

A técnica de ToF, permite a análise de massas através da razão massa/carga de um componente de uma amostra, sedo caracterizado por medida do tempo de voo dos íons dentro de um tupo a vácuo submetido a um campo magnético. Diferentes componentes diferem em massa e por tanto diferem em energia cinética

dentro desse campo, permitindo assim sua separação e determinação (Figura 1).

A espectrometria de massa MALDI-TOF é classicamente usada para análise de proteínas. A Figura 2 descreve um exemplo de uma proteína submetida à técnica de espectrometria de massa. A abscissa (eixo x) demonstra os valores da razão massa/carga da amostra enquanto a ordenada (eixo y) demonstra sua intensidade. Visivelmente podemos distinguir os pontos, picos, em que a intensidade se destaca do restante da amostra.

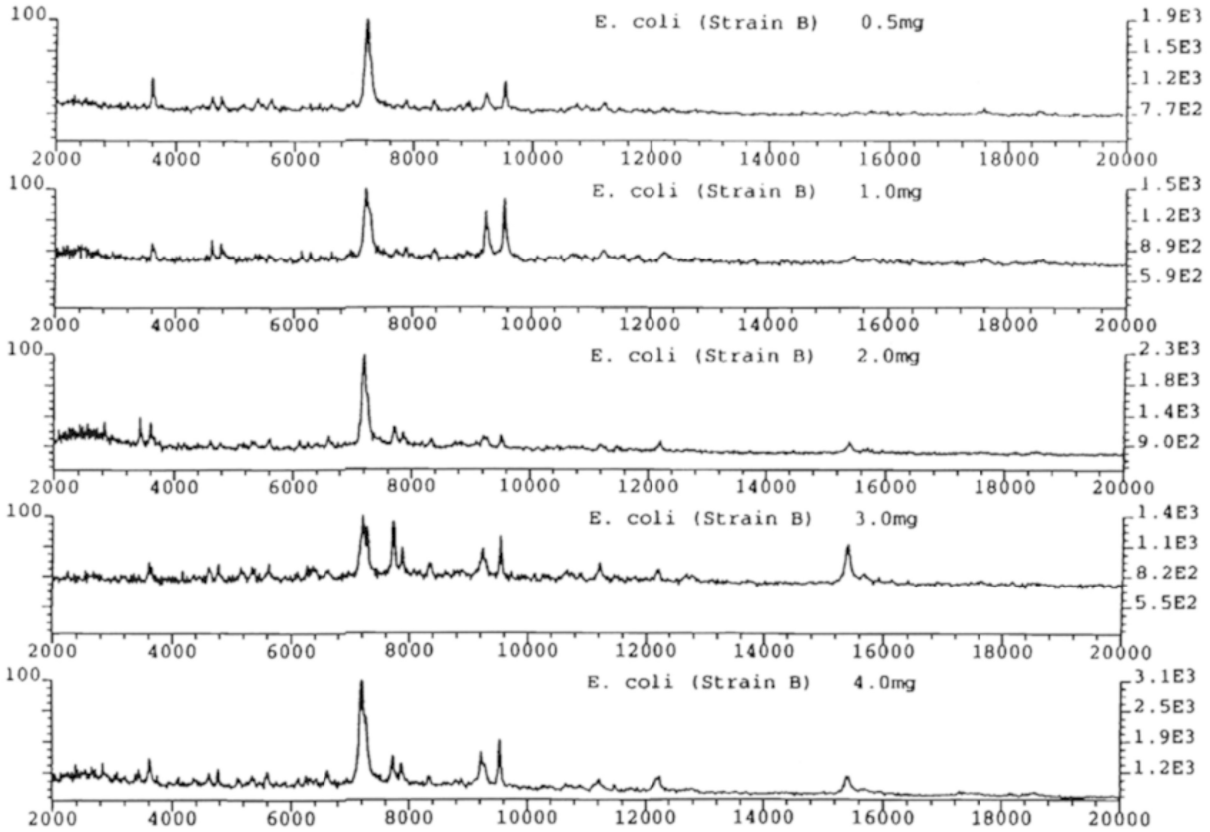
Figura 2: Horse Heart Myoglobin (HHM)



FONTE: Parsons, 2016

Mais recentemente, a espectrometria de massa MALDI-TOF tem sido aplicada para identificação de microrganismos. A Figura 3 descreve várias amostras de *Escherichia coli* submetidas ao mesmo processo de espectrometria de massa. Como visto nessa última, as amostras podem apresentar intensidades de picos distintas entre si, devido a diversos fatores. Entretanto a posição dos picos mantém-se igual. Isso permite, de certa forma, a criar uma assinatura proteômica, que logo pode ser utilizada para identificação de amostras.

Figura 3: Espectros de massa de amostras de Escherichia coli.



FONTE: Welham, 1998

2.3 APLICAÇÃO DA ESPECTROMETRIA DE MASSA NA IDENTIFICAÇÃO DE MICRORGANISMOS

A identificação de microrganismos é classicamente feita através de testes morfológicos (ex., formato de colônia), fenotípicos (ex., coloração Gram) e genéticos (ex., amplificação e sequenciamento de marcadores genéticos) (Santos; Hildenbrand; Schug, 2016).

Em anos recentes, a análise molecular, principalmente baseada na amplificação e sequenciamento de genes marcadores (ex., gene para 16S rRNA) tem se tornado técnicas importantes e se mostrado mais confiáveis e de maior resolução para a identificação de microrganismos.

Outra técnica contemporânea notável é a espectrometria de massa (MS). Principalmente o uso de espectrometria de massa MALDI-TOF tem se mostrado uma técnica molecular extremamente útil para identificação de microrganismos de forma confiável e rápida. Apesar do alto custo inicial para aquisição do

espectrômetro de massa, a técnica de MALDI-TOF MS permite a identificação de bactérias em poucos minutos, com um baixo custo por amostra, quando comparada a outras técnicas convencionais (Santos; Hildenbrand; Schug, 2016). De acordo com Tan (2012) o uso de MALDI-TOF para identificação de microrganismos gera uma redução de custos em até 10 vezes e uma redução do tempo de resultado em até 1,45 dias dos procedimentos padrões adotados no laboratório de microbiologia clínica do hospital “Johns Hopkins Hospital” (Estados Unidos da América).

Para a análise de bactérias por MALDI-TOF MS, normalmente os seguintes passos são cumpridos (Santos; Hildenbrand; Schug, 2016): i) a bactéria é geralmente crescida em meio de cultura sólido, produzindo colônias; ii) uma colônia é removida do meio de cultivo e transferida diretamente para placa de amostras do espectrômetro de massa; iii) a amostra é sobreposta com uma solução matriz, que irá co-cristalizar com as amostras e lisar as células; iv) a placa de amostras é colocada no espectrômetro de massa e bombardeado por laser, convertendo moléculas da célula em íons em fase gasosa que serão separados e identificados de acordo com sua razão massa/carga. O espectrômetro de massa fornece um padrão espectral característico para o microrganismo em análise, de modo a criar uma assinatura do mesmo.

A técnica de MALDI-TOF para identificação de microrganismos é principalmente utilizada na área clínica (Santos; Hildenbrand; Schug, 2016). A importância da rápida identificação de microrganismos envolvidos em infecções humanas e, consequentemente, a aplicação da terapia adequada, é inquestionável (Santos; Hildenbrand; Schug, 2016). Entretanto, a técnica de MALDI-TOF pode também fornecer uma significativa contribuição na microbiologia ambiental (Santos; Hildenbrand; Schug, 2016). Várias plataformas comerciais estão disponíveis para a identificação de microrganismos (Rahi; Prakash; Shouche, 2016), incluindo hardware e software, principalmente na área clínica: Bruker-Biotyper (Bruker Daltonics, Bremen, Alemanha), Axima Assurance (Shimadzu, Kyoto, Japão), SARAMIS AnagnosTec (AnagnosTec GmbH, Potsdam, Alemanha) e Andromas (Andromas SAS, Paris, França).

A análise dos espectros de massa para identificação de microrganismos são realizadas, principalmente, a partir de três abordagens gerais; i) identificação de

proteínas/peptídeos biomarcadoras específicas, indicativas de genótipo, propriedades fenotípicas ou taxa específicos; ii) pesquisas em banco de dados de proteína para agrupar ou identificar microrganismos; iii) comparações de múltiplos espectros para agrupamento taxonômico (Williams et al., 2003) e identificação de redundância em conjuntos de dados de organismos desconhecidos (Stets et al., 2013).

A identificação de microrganismos por espectrometria de massa tem aplicações potenciais em um grande número de áreas, dentre as quais: diagnóstica médica, biossegurança, monitoramento ambiental, agricultura, controle de qualidade de alimentos, segurança ocupacional e caracterização de cultura (Demirev; Fenselau, 2008).

Por outro lado, a técnica possui limitações, sendo as principais: pouca cobertura de bancos de dados, variações entre instrumentos, dependência de protocolo e resolução variável para diferenciação entre estirpes de bactérias. A solução para estas limitações virá do desenvolvimento de mais pesquisa na área (Wunschel et al., 2005). Por exemplo, o número limitado de bancos de dados contendo poucos organismos que não sejam de interesse clínico, resulta em baixa porcentagem de identificação (43 a 65%) para microrganismos isolados do solo, água e outros ambientes (Rahi; Prakash; Shouche, 2016).

2.4 PROGRAMAS E BANCOS DE DADOS PARA ANÁLISE DE ESPECTROS DE MASSA

A identificação de microrganismos através da análise por MALDI-TOF MS é dependente de bancos de dados contendo espectros de massa de estirpes referências bem conhecidas e caracterizadas (Santos; Hildenbrand; Schug, 2016). A identificação pode ser feita por comparação entre o perfil proteico desconhecido contra um banco de dados de perfis de referência ou então por análise de agrupamento do perfil desconhecido com perfis de bactérias conhecidas (Santos; Hildenbrand; Schug, 2016). A primeira etapa destas análises requer a comparação direta entre cada par de espectros no conjunto de dados, para identificação de picos comuns e estabelecimento do grau de similaridade ente os espectros. Correlações para as posições e intensidades dos picos entre espectros das amostras experimentais e aqueles compondo o banco de dados são usados para gerar uma

pontuação (*score*), que representa um nível de confiança que o isolado desconhecido é um representante do microrganismo candidato incluído no banco de dados (Santos; Hildenbrand; Schug, 2016). Atualmente, dois principais bancos de dados estão disponíveis: Bruker BioTyper (Bruker Daltonics, Inc.) e SARAMIS (bioMérieux). Os bancos de dados utilizam o Bruker *Main Spectrum analysis* (MSP) e o bioMérieux SuperSpectrum, respectivamente. Estas abordagens diferem no algoritmo usado para a identificação dos microrganismos (Santos; Hildenbrand; Schug, 2016). O MSP consiste de uma coleção de espectros de referência obtidos a partir de uma única estirpe referência, enquanto o bioMérieux SuperSpectrum consiste de um espectro obtido a partir de várias estirpes clínicas e referências crescidas sob diferentes condições (Rahi; Prakash; Shouche, 2016)(Santos; Hildenbrand; Schug, 2016).

Até recentemente, os desenvolvedores de sistemas para identificação de microrganismos por espectrometria de massa MALDI-TOF tem como principal foco a área clínica e a principal limitação atribuída ao uso da técnica na ecologia microbiana tem sido atribuída a falta de espectros de referência nos bancos de dados associados aos instrumentos (Rahi; Prakash; Shouche, 2016). Outra dificuldade é a incapacidade de utilização de dados entre diferentes plataformas. A solução dada tem sido a criação de bancos de dados locais (Rahi; Prakash; Shouche, 2016).

2.5 SPECLUST

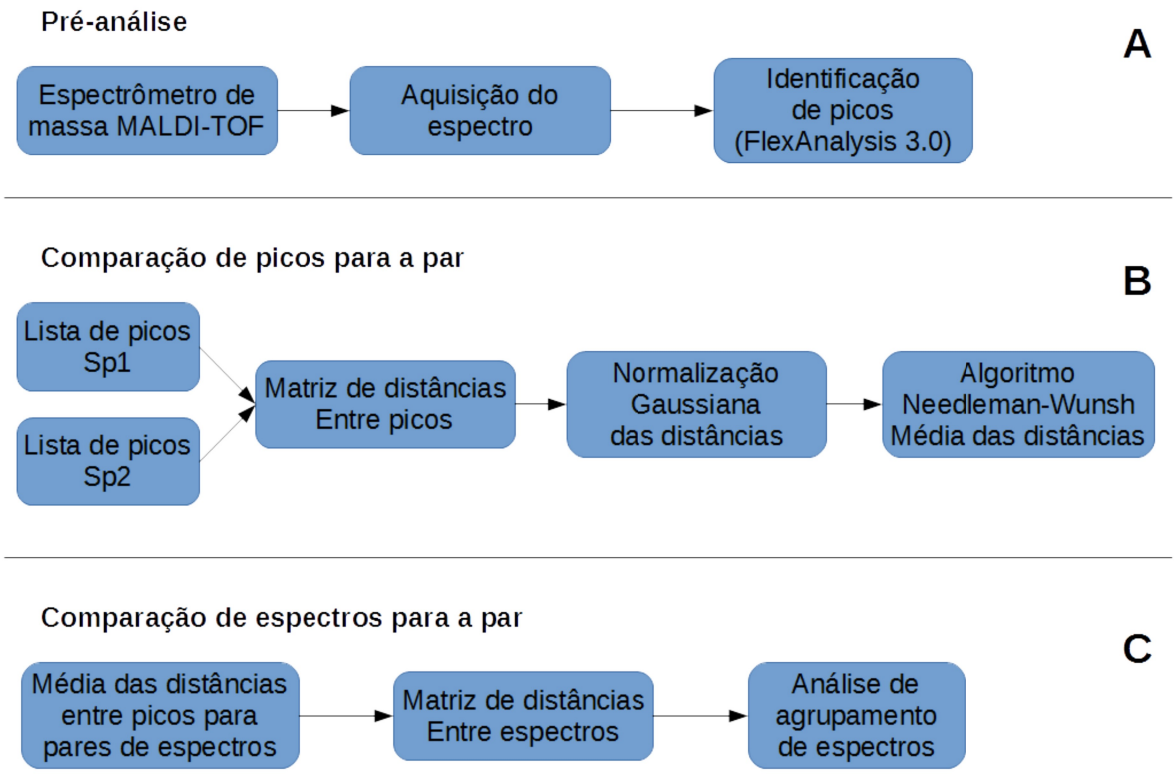
O programa foi inicialmente desenvolvido para varredura de grandes conjuntos de dados de espectros de massa MALDI-TOF para identificação de isoformas de proteínas separadas em géis de eletroforese bidimensional (Alm et al., 2006).

O programa Speclust compara inicialmente dois espectros de massa MALDI-TOF, através da lista de picos previamente identificados para cada um dos espectros. A comparação gera uma matriz de distância para todos os pares de picos entre os dois espectros. Em seguida, a esta matriz, é aplicada uma normalização Gaussiana e, posteriormente, calculada a média da distância entre os dois espectros, usando um algoritmo semelhante ao Needleman-Wunsch usado em alinhamento de sequências de nucleotídios e aminoácidos (Figura 4). A partir desta

comparação para a par entre espectros de massa, o programa permite a criação de uma segunda matriz de similaridade entre três ou mais espectros de massa e que pode ser utilizada para o posterior agrupamento entre espectros (Figura 4).

O processo de agrupamento realizado pelo Speclust utiliza o algoritmo sugerido por Ward (1963). O método inicia designando cada lista de picos ao seu próprio agrupamento e calculando uma distância entre cada par da lista de picos. O par de picos mais próximo é encontrado e unido em um novo agrupamento. Distâncias entre o novo agrupamento e cada um dos agrupamentos anteriores são calculadas. A busca pelo par mais próximo, união do par e cálculo das novas distâncias são repetidas até que haja um único agrupamento. O agrupamento é realizado a partir do método de ligação por média (do inglês, *average linkage*), onde a distância entre dois agrupamentos é calculada como a média das distâncias a partir de cada lista de picos em um agrupamento para cada lista de picos no outro agrupamento.

Figura 4: Fluxograma de análise do programa Speclust



FONTE: O autor (2017)

3 MATERIAIS E MÉTODOS

3.1 EQUIPAMENTOS

O programa foi desenvolvido e testado em computador com processador Intel i7-4500u, 16GB 1600MHz DDR3 RAM, Kingston SATA3 480 GB SSD HD e SO Windows 10 Home Single Language.

3.2 LINGUAGEM DE PROGRAMAÇÃO

A linguagem escolhida para implementar o projeto foi a JavaSE 8u92. Suas capacidades natas favorecem a conclusão dos objetivos específicos deste projeto, considerando que Java é notoriamente conhecido por permitir que múltiplas plataformas executem o mesmo arquivo executável sem a necessidade de nova compilação.

3.3 DEPENDÊNCIAS

Duas bibliotecas Java foram utilizadas e fazem parte do programa desenvolvido:

1. JFreechart 1.0.19. Biblioteca para geração de gráficos e mantida por David Gilbert; utilizada para facilitar a criação e exibição de gráficos a partir dos dados coletados e analisados; código aberto; disponível em <www.jfree.org>.
2. Ini4j 0.5.4. Uma API Java para manipulação de arquivos no formato “.ini” do Windows; utilizada para facilitar a manipulação, leitura e escrita de arquivos; código aberto; disponível em <ini4j.sourceforge.net>.

3.4 AMBIENTE DE DESENVOLVIMENTO

O Eclipse Mars.2, utilizado neste projeto, é um dos principais ambientes integrados de desenvolvimento, disponibilizando IDE (do inglês, *Integrated Development Environment*) e plataformas para linguagens de programação, como Java, C/C++, Javascript e PHP. Apresenta baixo uso de memória e processamento e é uma das principais ferramentas de desenvolvimento de código livre disponível. Disponível em <eclipse.org>.

3.5 DADOS DE ENTRADA

Os dados, alvo de análise, utilizados no projeto, como forma de teste e

validação do aplicativo são os arquivos XML com a lista de picos de espectros. Esses arquivos foram fornecidos pelo Setor de Bioquímica da Universidade Federal do Paraná, após serem obtidos pelo processamento dos dados brutos do espectrômetro de massa Bruker Daltonics Autoflex II, pelo software proprietário Bruker Daltonics FlexAnalysis 3.0.

3.6 ARQUIVO DE ESPECTRO DE MASSA EM XML

Diversos arquivos são gerados automaticamente pelo espectrômetro de massa MALDI-TOF, modelo Autoflex II (Bruker Daltonics, Bremen, Alemanha), gerenciado pelo programa FlexControl 3.0 (Bruker Daltonics, Bremen, Alemanha). O arquivo bruto é posteriormente convertido em um arquivo contendo a lista de picos m/z identificados através do programa FlexAnalysis 3.0 (Bruker-Daltonics, Bremen, Alemanha). Os arquivos, para cada amostra analisada, são armazenados em uma série de diretórios hierárquicos (Figura 5). Os diretórios são mostrados em negrito e seguidos de “/”. O arquivo “peaklist.xml” contém informações sobre o espectro de massa obtido para a amostra e é usado no programa JMSA.

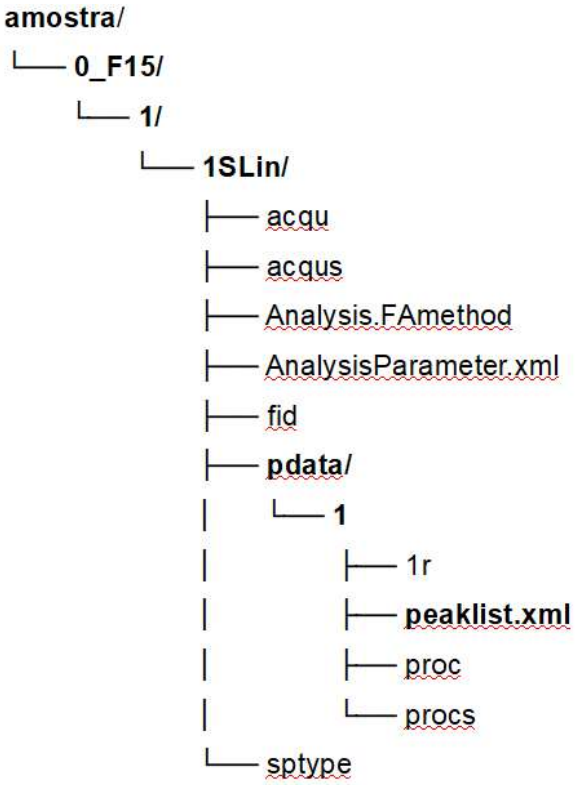
De forma resumida os arquivos XML contendo a lista de picos dos espectros de massa são obtidos nesta sequência:

1. MALDI-TOF AutoFlex II (Bruker Daltonics, Bremen, Alemanha) – aparelho espectrômetro de massa para aquisição dos dados brutos de relação m/z a partir das amostras

2. FlexControl 3.0 (Bruker Daltonics, Bremen, Alemanha) – software para gerenciamento do espectrômetro de massa e armazenamento dos dados brutos

3. FlexAnalysis 3.0 (Bruker-Daltonics, Bremen, Alemanha) – software para identificação de picos m/z a partir do espectro bruto e armazenamento da lista de picos identificados em formato XML.

Figura 5: Exemplo de hierarquia de pastas a ser percorrida recursivamente pelo programa.



FONTE: O autor (2017)

Exemplo de arquivo “peaklist.xml”, em texto formatado XML, contendo informações sobre um único pico (Figura 6). Neste tipo de arquivo de texto estruturado, a informação é incluída entre marcadores ou etiquetas (do inglês, *tags*). A informação pode ser aninhada entre marcadores de forma hierárquica. Por exemplo, os marcadores “<pk>” e </pk>” incluem informações a respeito de um pico no espectro de massa e podem incluir outros marcadores com informações específicas, como, por exemplo, “<area>” e ”</area>”, contendo a área do pico, ou “<mass>” e “</mass>”, contendo a razão m/z para o pico. O marcador “<pklist>” inclui informações gerais sobre a aquisição do espectro de massa, incluídas na forma de marcador=”valor”, por exemplo indicando a data (date=”2013-06-19T11:40:42.265+01:00”) ou o número de tiros (shots=”1000”) na análise da amostra.

Figura 6: Extração parcial de arquivo XML gerado pelo software FlexAnalysis 3.0 contendo informações sobre os picos m/z detectados em um espectro de massa MALDI-TOF

```
<?xml version="1.0"?>
<pklist
    version="1.0"
    creator="flexAnalysis 3.0.92.0"
    shots="1000" date="2013-06-19T11:40:42.265+01:00"
    spectrumid="AamazonenseY2r1_0A1xml"
    spectrumid2="1c1ff4bc-3ff0-4464-9950-15d518f9428c">
<pk>
    <absi>334.5000000000000</absi>
    <area>2064.232230821399</area>
    <ind>544.9795371427754</ind>
    <lind>494.0000000000000</lind>
    <mass>3700.407367382949</mass>
    <meth>2</meth>
    <reso>866.1307932591151</reso>
    <rind>571.0000000000000</rind>
    <s2n>8.073897924132604</s2n>
    <type>0</type>
</pk>
```

FONTE: O autor (2017)

3.7 DISTÂNCIA EUCLIDIANA

A Distância Euclidiana é a medida mais conhecida para indicar a dissimilaridade para dados numéricos quantitativos e é definida como a distância geométrica no espaço multidimensional, sendo calculada por meio da fórmula:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Dessa forma é possível determinar, num espaço de ***n*** dimensões o quão distantes são dados dois pontos nesse espaço, e por consequência ter uma medida de divergência entre dois ou mais pontos (CRUZ; CARNEIRO, 2003).

Podemos assim determinar os valores de massa/carga (m/z) como uma dimensão na distância euclidiana.

3.8 DISTRIBUIÇÃO NORMAL

Para entender a distribuição normal é preciso entender a função gaussiana

da qual a distribuição é derivada. A função gaussiana é descrita por

$$f(x)=a*\exp\left(\frac{-(x-b)^2}{2c^2}\right)$$, sendo que:

- *a* indica o pico da curva
- *b* é a posição central da curva
- *c* indica o desvio padrão da curva

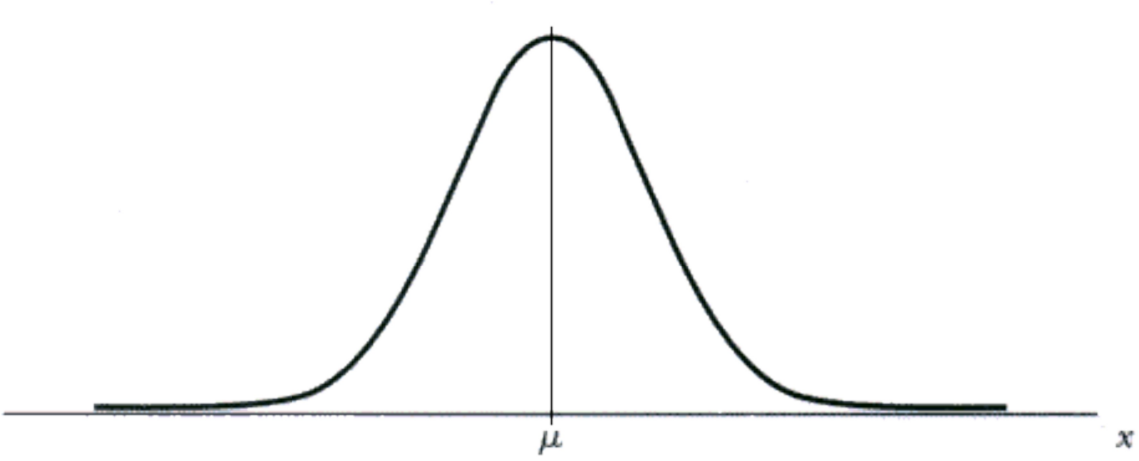
O uso dessa função para a normalização de dados de forma a distribuí-los numa gama contínua de possibilidades é chamada de distribuição normal.

Primeiramente utilizada pelo matemático francês Abraham De Moivre, a distribuição normal, também conhecida por distribuição gaussiana é uma distribuição contínua apresentada pela fórmula:

$$f(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}*\left(\frac{x-\mu}{\sigma}\right)^2\right), x\in(-\infty,\infty)$$

Tal distribuição é representada simetricamente em forma de sino ao redor do centro estabelecido pela média dos dados. Semelhantemente ao mostrado pela Figura 7.

Figura 7: Distribuição normal



FONTE: <http://fm.usp.br/dim/diststat/index.php> Acesso em 29 de janeiro de 2017

Esse gráfico descreve em sua área a probabilidade de ocorrência de diversos fenômenos aleatórios que ocorrem na natureza, indústria e pesquisa.

Dentre as características mais notáveis que compõem a curva da distribuição normal destaca-se que ela é simétrica em relação ao centro e esse centro coincide

com o máximo da função sendo que $f(x)$ possui máximo em $(x=\mu)$. Vale citar que a área total compreendida pela curva é igual a 1, logo 100% de probabilidade.

Quando se utiliza o modelo normal, o valor da média indica o centro da distribuição, e o desvio padrão mede a dispersão do conjunto, indicando a variabilidade em relação ao centro.

O cálculo da área desta região exige recursos do cálculo diferencial integral. Entretanto, tal cálculo pode ser simplificado quando utilizamos valores tabelados, que permitem encontrar facilmente os valores de probabilidades desejados. Chamamos assim de distribuição normal padronizada.

Uma distribuição normal é dita padronizada quando $\mu=0$ e $\sigma=1$.

Fazendo com que sua função seja reduzida para $f(x)=\frac{1}{\sqrt{2\pi}}\exp(\frac{-x^2}{2})$.

3.9 ALGORITMO NEEDLEMAN-WUNSCH

O algoritmo Needleman-Wunsch (Needleman; Wunsch, 1970) é classicamente utilizado na bioinformática para alinhamento de sequências de nucleotídeos ou aminoácidos.

O funcionamento do algoritmo depende do desenvolvimento de uma matriz de *traceback*, feita a partir de sistema de pontuação. O sistema de pontuação deve abranger três casos determinados pelo usuário do algoritmo: *match* (pareamento correto)*, mismatch* (pareamento incorreto) *e gap* (inclusão de lacunas).

Para fins de exemplificação utilizaremos o seguinte sistema de pontuação:

Tabela 2: Sistema de pontuação

Caso	**Pontuação**
match	+1
mismatch	-1
gap	-2

FONTE: O autor (2017)

Em seguida uma matriz de similaridade é feita utilizando a projeção de uma sequência contra outra incluindo uma linha inicial em branco e uma coluna inicial também em branco. Na linha em branco adicionada, preenche-se a primeira casa com 0 e para cada coluna adiciona-se uma pontuação de *gap*. Na coluna em branco,

a primeira casa já deve estar preenchida com 0, então para cada linha adiciona-se uma pontuação de *gap*.

O restante das casas da matriz é preenchida segundo a soma da maior pontuação possível, considerando os 3 casos previstos, de tal forma que *match* e *mismatch* somam-se a casa diretamente na diagonal acima e a esquerda, enquanto o *gap* soma-se a casa à esquerda ou acima.

Tabela 3: Matriz de pontuação

Seq.		**A**	**C**	**T**	**G**	**A**	**T**	**T**	**C**	**A**
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	1	-1	-3	-5	-7	-9	-11	-13	-15
C	-4	-1	2	0	-2	-4	-6	-8	-10	-12
G	-6	-3	0	1	1	-1	-3	-5	-7	-9
C	-8	-5	-2	-1	0	0	-2	-4	-4	-6
A	-10	-7	-4	-3	-2	1	-1	-3	-5	-3
T	-12	-9	-6	-3	-4	-1	2	0	-2	-4
C	-14	-11	-8	-5	-4	-3	0	1	1	-1
A	-16	-13	-10	-7	-6	-3	-2	-1	0	2

FONTE: O autor (2017)

Por fim basta rastrear o caminho da maior pontuação para determinar as possibilidades de melhor alinhamento (etapa de *traceback*). Na tabela, as linhas verdes representam um *MATCH*, as amarelas representam um *MISMATCH,* e as vermelhas representam um *GAP*.

Tabela 4: Matriz de Traceback

Seq.		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	1	-1	-3	-5	-7	-9	-11	-13	-15
C	-4	-1	2	0	-2	-4	-6	-8	-10	-12
G	-6	-3	0	1	1	-1	-3	-5	-7	-9
C	-8	-5	-2	-1	0	0	-2	-4	-4	-6
A	-10	-7	-4	-3	-2	1	-1	-3	-5	-3
T	-12	-9	-6	-3	-4	-1	2	0	-2	-4
C	-14	-11	-8	-5	-4	-3	0	1	1	-1
A	-16	-13	-10	-7	-6	-3	-2	-1	0	2

FONTE: O autor (2017)

LEGENDA: *MATCH*; *MISMATCH*; *GAP*.

O alinhamento das sequências, dado o sistema de pontuação, resulta nas seguintes possibilidades:

Tabela 5: Melhores resultados de alinhamento

	Alinhamento 1	Alinhamento 2	Alinhamento 3
ACTGATTCA	**ACTGATTCA**	**ACTG-ATTCA**	**ACTGA-TTCA**
ACGCATCA	**ACGCAT-CA**	**ACG-CAT-CA**	**ACG-CAT-CA**

FONTE: O autor (2017)

4 RESULTADOS E DISCUSSÃO

Neste trabalho foi desenvolvido o programa JSMA – Java Mass Spectrometry Analyzer, para manipulação, visualização e análise de espectros de massa MALDI-TOF para identificação de microrganismos. O programa tem como principais características:

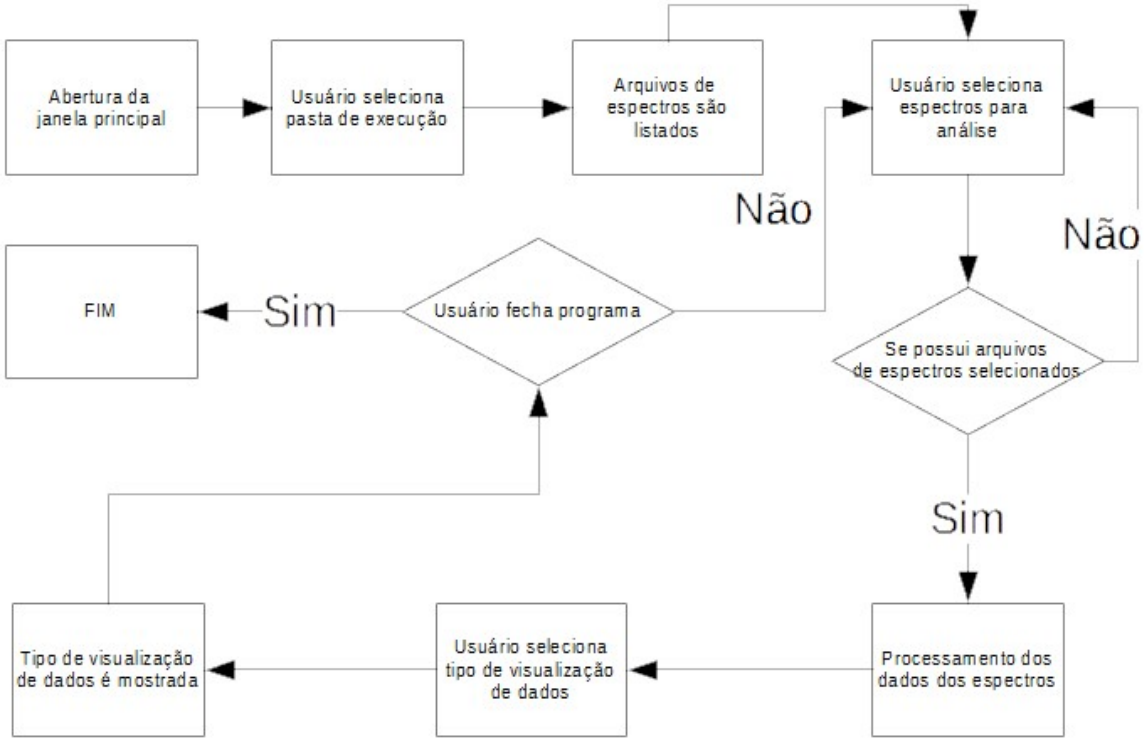
i. importar arquivos no formato XML pré-analisados para a identificação de picos através do programa FlexAnalysis 3.0 (Bruker Daltonics);
ii. armazenar informações correspondentes a análise e a amostra biológica no formato XML;
iii. comparar espectros de massa MALDI-TOF;
iv. visualizar e editar as informações relativas a análise de espectrometria de massa e em relação às amostras biológicas.
v. Programa de código aberto

JMSA foi desenvolvido utilizando a linguagem de programação Java SE JDK 8u92, de forma a favorecer seu uso em múltiplos sistemas operacionais. No produto foram utilizados as bibliotecas JFreechart para criar os gráficos que representam os espectros de massa e ini4j para manipulação de arquivos de configuração e para o armazenamento das informações sobre cada espectro.

Por ser um programa de código aberto, o custo da análise de dados de espectrometria de massa é reduzida.

A primeira entrada que o usuário oferece no aplicativo é o diretório raiz onde se encontram os arquivos XML a serem analisados, obtidos por pré-análise no programa Bruker Daltonics FlexAnalysis 3.0. Esse diretório é então percorrido recursivamente para listar os arquivos considerados válidos, por conterem os campos de informação utilizados pelo JMSA.

Figura 8: Fluxograma simplificado de execução do JMSA.



FONTE: O autor (2017)

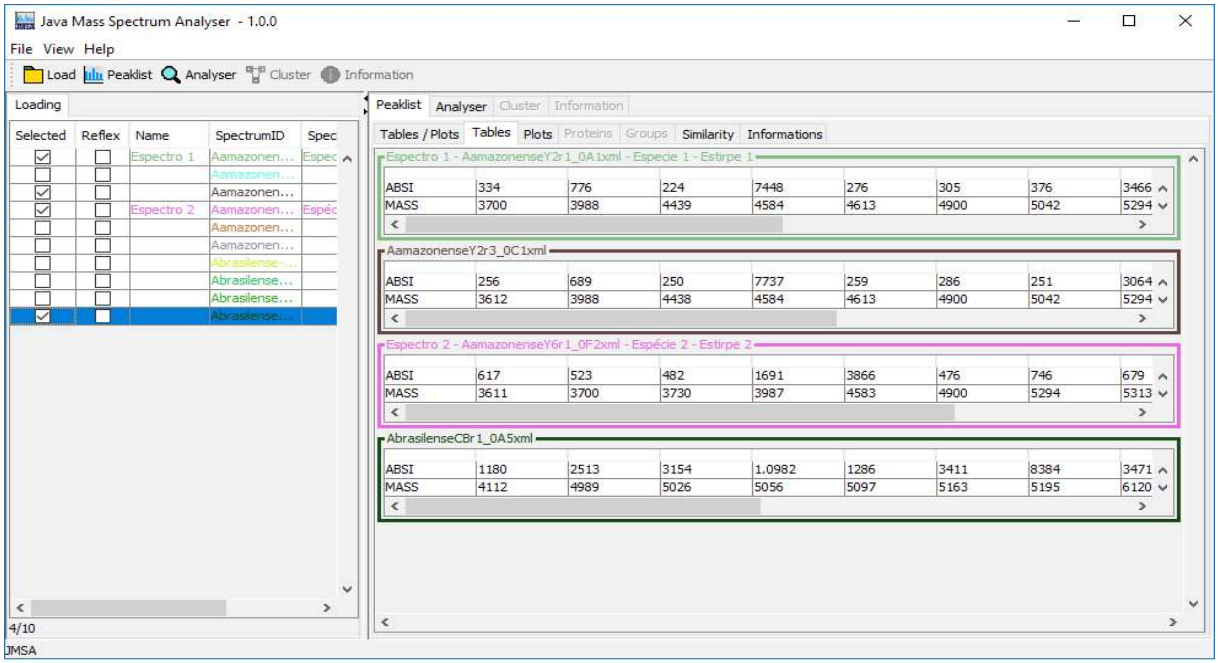
Uma vez que essa lista de espectros está carregada, o usuário pode selecionar um ou mais espectros para visualização, edição e análise. Os arquivos carregados são listados na janela à esquerda e cada espectro é identificado a partir de informações extraídas do arquivo XML. Cada espectro é primariamente identificado pelo campo “SpectrumID” do arquivo XML e criado automaticamente pelo programa FlexAnalysis 3.0. Outras informações podem ser inseridas pelo usuário e armazenadas em arquivo próprio, mostrando-as na listagem para facilitar a identificação dos espectros. Por exemplo, a lista poderia ser ordenada por organismos. O arquivo XML original é mantido íntegro e é gerado um novo arquivo formatado para armazenar as edições realizadas pelo usuário, no mesmo diretório do arquivo original e com a extensão “jsmainfo”, sendo:

1. Tabela de picos do espectro, com comparação lado a lado, visto em Figura 9.
2. Gráfico de picos do espectro, espectro único em tela cheia visto em Figura 10, múltiplos espectros em comparação lado a lado, visto em Figura 11.
3. Tabela de picos aninhada ao gráfico de picos, com comparação lado a lado,

visto em Figura 12.

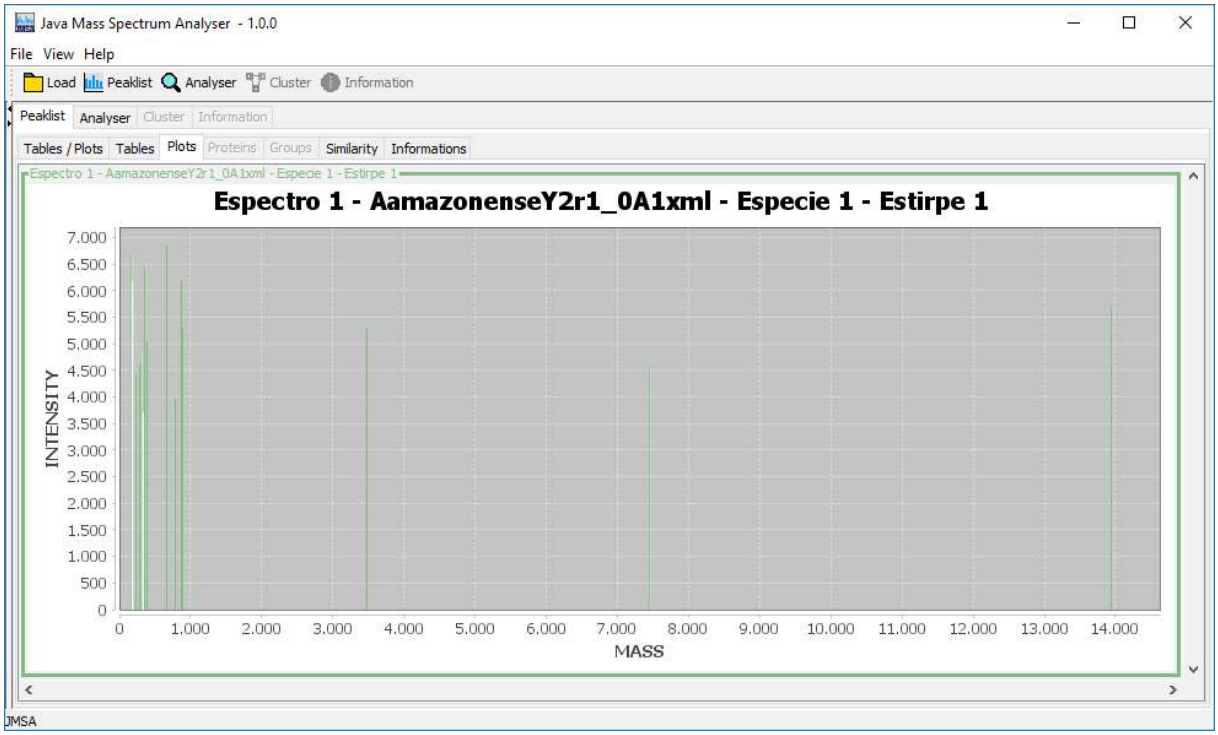
4. Tabela de similaridade entre espectros, visto em Figura 13.
5. Formulário de informações relacionadas ao espectro (Figura 14). As informações biológicas associadas a um espectro são armazenadas em um arquivo de texto estruturado, no mesmo diretório do arquivo XML com informações do espectro de massa e com extensão “jsmainfo” (Figura 15).

Figura 9: Comparação dos pontos de picos de espectros selecionados



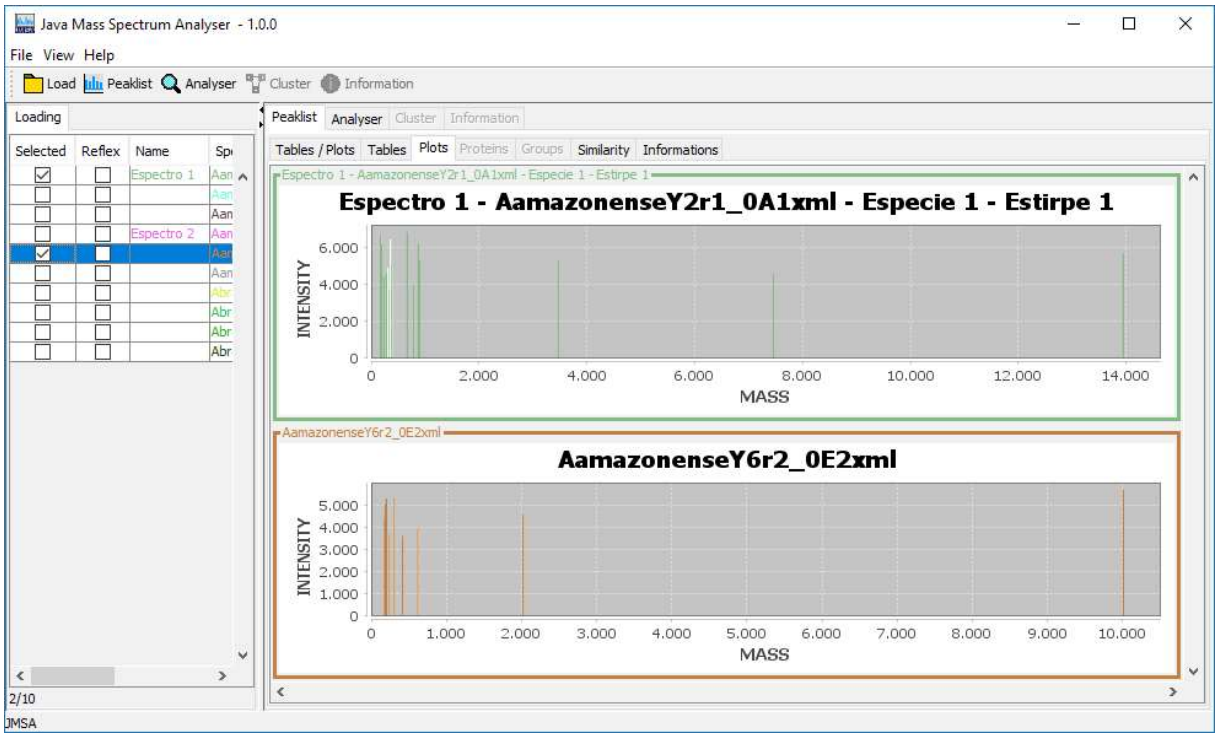
FONTE: O autor (2017)

Figura 10: Gráfico de picos de espetro, utilizando completamente o espaço de visualização.



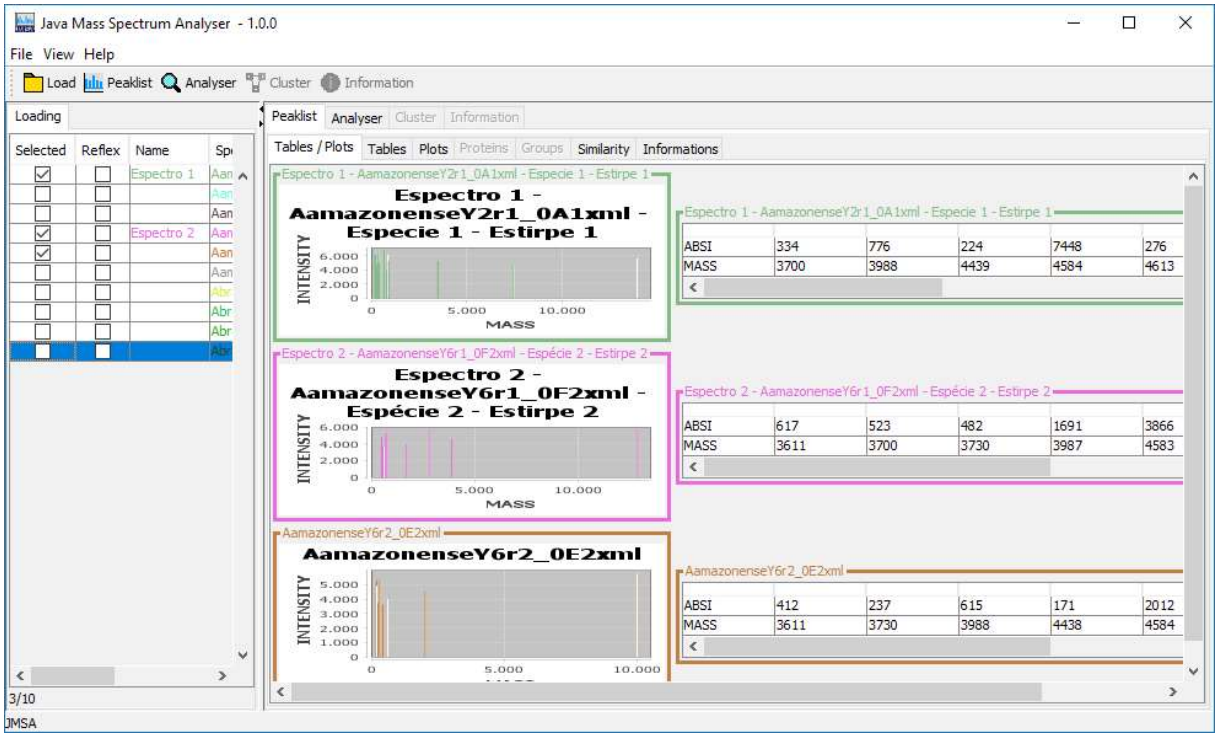
FONTE: O autor (2017)

Figura 11: Gráfico de picos de espetro, utilizando dois espectros selecionados que compartilham o espaço de visualização.



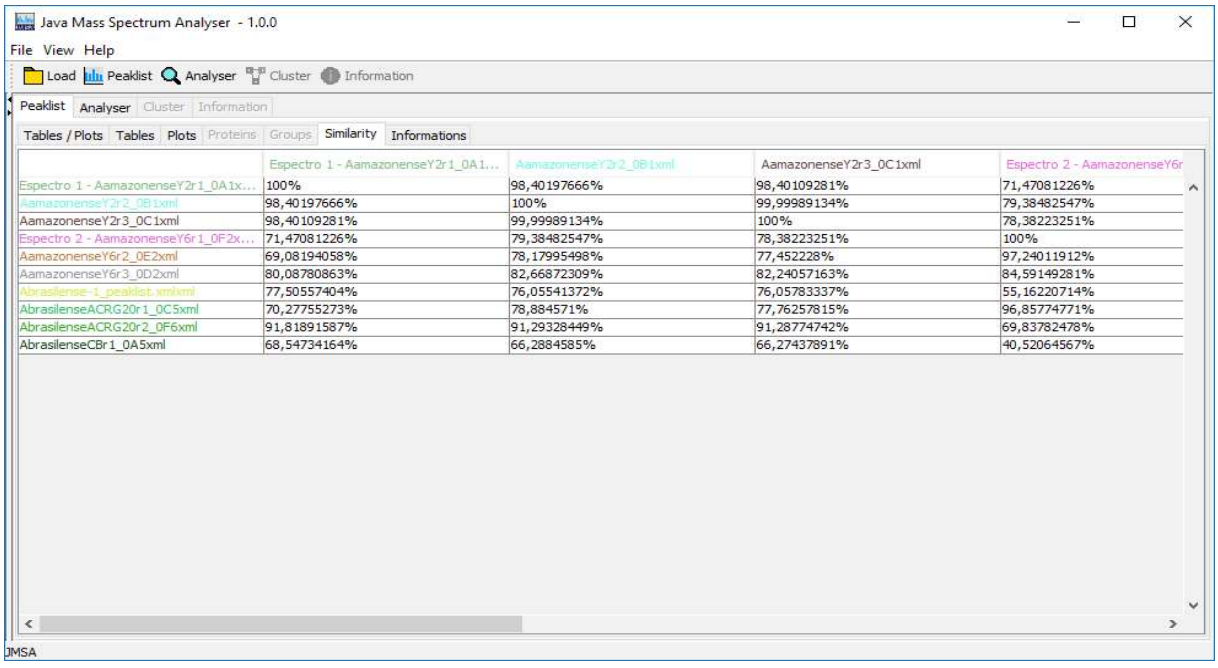
FONTE: O autor (2017)

Figura 12: Modo de visualização em que os espectros selecionados compartilham o espaço de visualização verticalmente entre si enquanto mostram seus gráficos de picos junto com suas tabelas de pontos de pico horizontalmente



FONTE: O autor (2017)

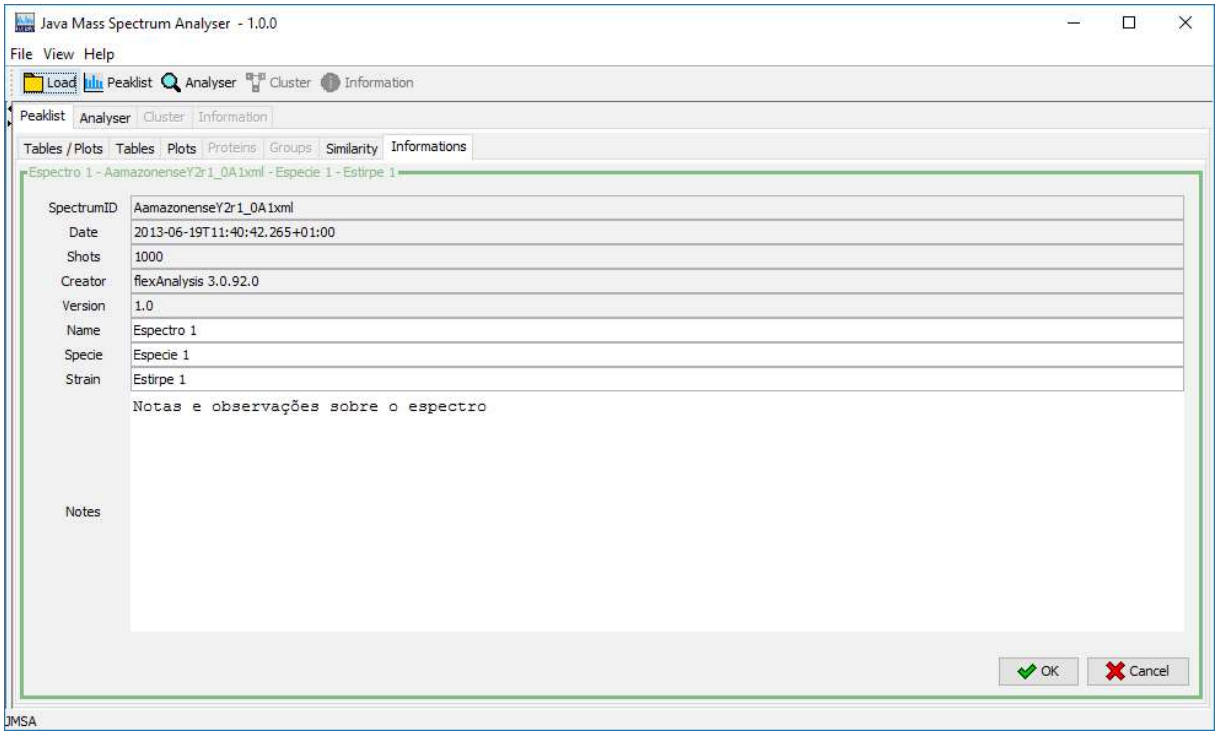
Figura 13: Tabela que mostra a porcentagem de similaridade entre todos os espectros selecionados.



	Espectro 1 - AamazonenseY2r1_0A1...	AamazonenseY2r2_0B1xml	AamazonenseY2r3_0C1xml	Espectro 2 - AamazonenseY6r
Espectro 1 - AamazonenseY2r1_0A1x...	100%	98,40197666%	98,40109281%	71,47081226%
AamazonenseY2r2_0B1xml	98,40197666%	100%	99,99989134%	79,38482547%
AamazonenseY2r3_0C1xml	98,40109281%	99,99989134%	100%	78,38223251%
Espectro 2 - AamazonenseY6r1_0F2x...	71,47081226%	79,38482547%	78,38223251%	100%
AamazonenseY6r2_0E2xml	69,08194058%	78,17995498%	77,452228%	97,24011912%
AamazonenseY6r3_0D2xml	80,08780863%	82,66872309%	82,24057163%	84,59149281%
Abrasilense-1_peaklist.xmlxml	77,50557404%	76,05541372%	76,05783337%	55,16220714%
AbrasilenseACRG20r1_0C5xml	70,27755273%	78,884571%	77,76257815%	96,85774771%
AbrasilenseACRG20r2_0F6xml	91,81891587%	91,29328449%	91,28774742%	69,83782478%
AbrasilenseCBr1_0A5xml	68,54734164%	66,2884585%	66,27437891%	40,52064567%

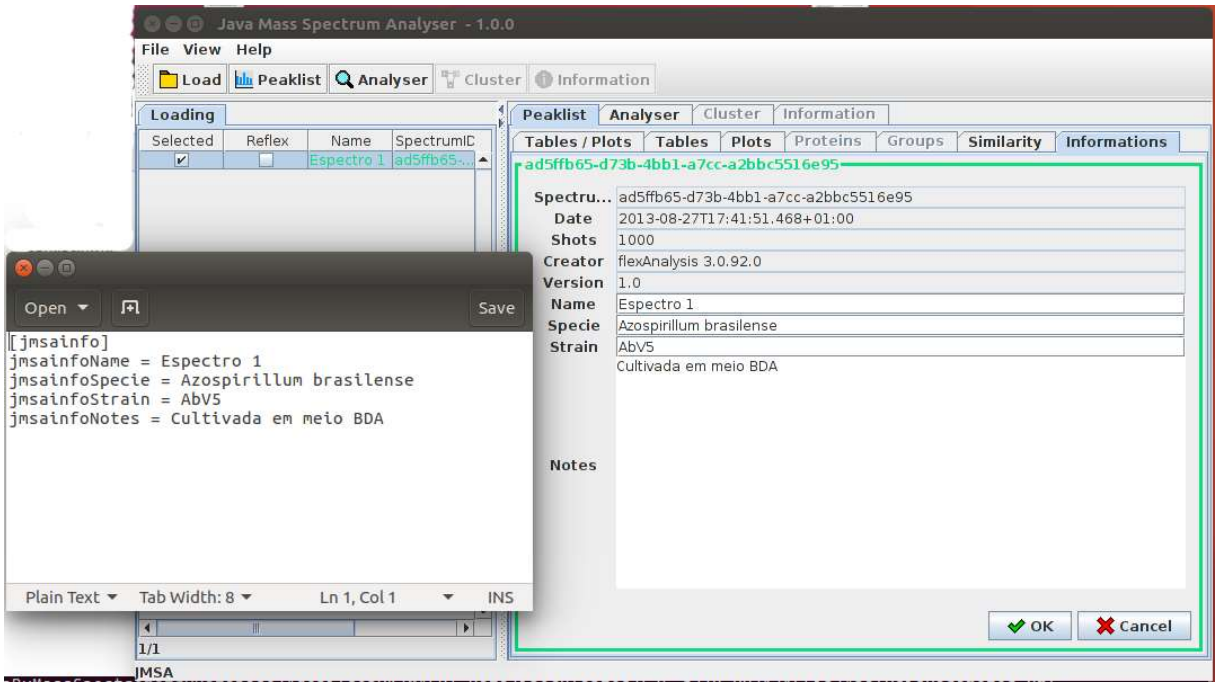
FONTE: O autor (2017)

Figura 14: Formulário com as informações pertinentes a um espectro selecionado.



FONTE: O autor (2017)

Figura 15: Entrada de informações biológicas no programa JSMA e a estrutura do arquivo de texto "jsmainfo" que armazena as informações.

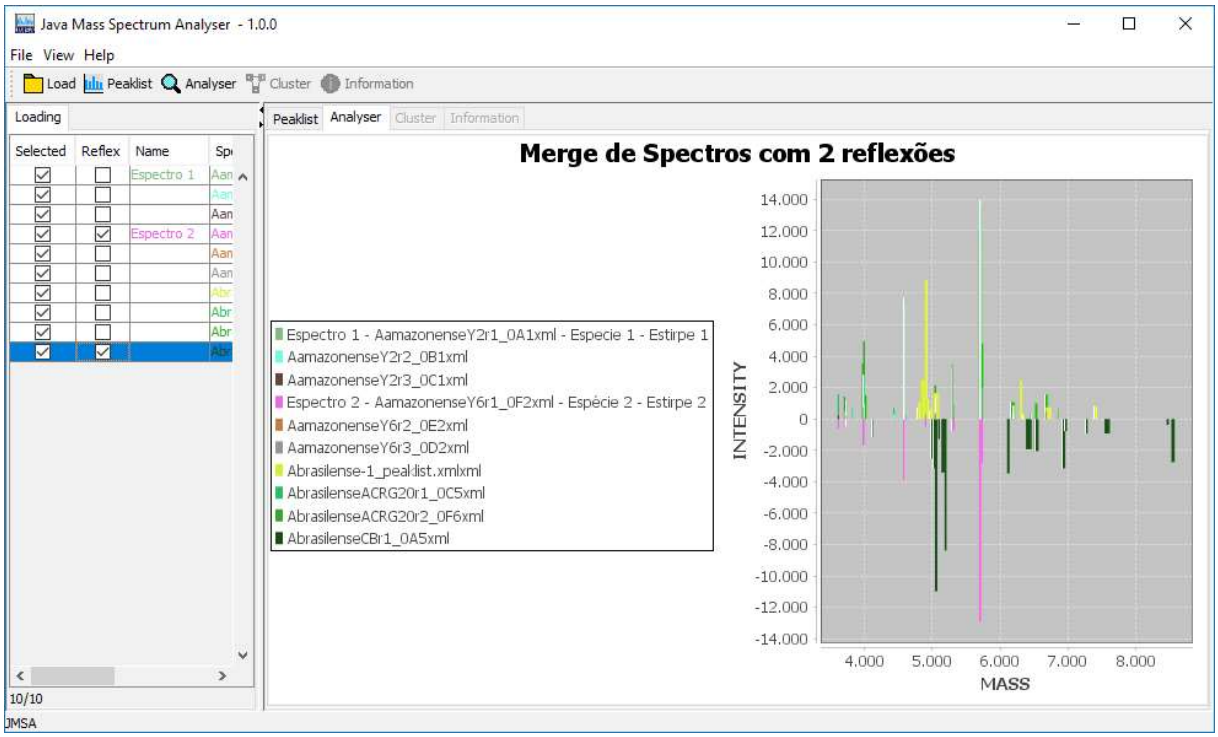


FONTE: O autor (2017)

A função “SuperSpectro” do software JSMA (Figura 16) permite agrupar

diferentes espectros de massa em “um único”. Desta forma, esta função contempla a variação inerente à técnica de espectrometria de massa MALDI-TOF para identificação de microrganismos. Por exemplo, para um mesmo organismo, toda a variação observada em função de replicadas técnicas e biológicas, meios de cultivo, métodos de extração, etc. podem ser representadas em um único superespectro. Desta forma, a confiabilidade da comparação entre organismos pode ser aumentada quando feita através de superespectros.

Figura 16: SuperSpectro. Fusão entre os dados de picos de todos os espectros selecionados. Nota-se que dois dos espectros selecionados estão marcados para refletir seus graficos.



FONTE: O autor (2017)

4.1 IDENTIFICAÇÃO DE MICRORGANISMOS COM O SOFTWARE JSMA

O software JSMA permite a comparação entre vários espectros e seu agrupamento de acordo com a similaridade. A Figura 17 mostra a capacidade da técnica de espectrometria de massa MALDI-TOF e da análise de comparação implementada no software JSMA de agrupar bactérias da mesma espécie. A figura mostra que os espectros pertencentes a uma mesma espécie possuem os valores mais altos de similaridade. No caso do gênero *Azospirillum*, é possível observar também que os espectros de massa correspondentes às espécies *A. brasilense* e *A.*

amazonense foram agrupadas corretamente.

Figura 17: Comparação entre vários espectros de massa de diferentes organismos no software JSMA. As porcentagens indicam o grau de similaridade entre dois espectros de massa a partir da comparação dos picos; em verde, a similaridade entre o espectro de massa com ele próprio; em vermelho, a similaridade dos espectros de massa indicado nas colunas com os dois mais próximos.

	Azospirillum amazonense	Azospirillum brasilense	Aeromonas hidrofila	Escherichia coli
Azospirillum amazonense	100%	40,79582569%	48,89154027%	70,75325837%
Azospirillum amazonense	99,99989134%	40,10187796%	48,59301456%	72,53440241%
Azospirillum amazonense	78,38223251%	28,61757225%	32,35077275%	61,41484786%
Escherichia coli	70,75325837%	60,15243565%	68,62999937%	100%
Escherichia coli	70,54234397%	63,10220803%	68,97659346%	99,04071767%
Azospirillum brasilense	63,86678655%	87,84970797%	76,65359082%	73,32707713%
Escherichia coli	62,86645219%	68,86384081%	83,09436197%	88,09990623%
Aeromonas hidrofila	52,58316648%	72,77273787%	96,04881241%	73,39726066%
Azospirillum brasilense	49,33599926%	94,05219627%	73,31825878%	57,72146849%
Aeromonas hidrofila	48,89154027%	72,46170722%	100%	68,62999937%
Aeromonas hidrofila	42,95508672%	77,78305351%	96,07079966%	65,40538472%
Azospirillum brasilense	40,79582569%	100%	72,46170722%	60,15243565%

FONTE: O autor (2017)

Esta análise pode ser estendida para a identificação de espectros de massa obtidos a partir de bactérias cultivadas desconhecidas, quando comparado a um banco de dados de espectros de massa de bactérias referência. Esta situação é ilustrada na Figura 18, onde um espectro de massa de organismo previamente identificado, obtido a partir da bactéria *Azospirillum amazonense* Y2, foi adicionado à análise, com o grupo de espectros de massa da análise anterior (Figura 17). É possível observar que a similaridade do espectro desconhecido é maior com espectros de *A. amazonense*, sugerindo a identificação ou proximidade taxonômica para o organismo desconhecido. Nesta simulação, o espectro de massa foi submetido a comparação de similaridade com outros espectros de massa, tal como se fosse desconhecido. Tendo a tabela de similaridade apontado os espectros da espécie *Azospirillum amazonense como maiores similaridades para com este mostra, portanto, que* o programa foi capaz de identificar a amostra corretamente em nível de espécie.

Figura 18: Comparação entre um espectro de massa desconhecido com espectros de massa referência para identificação de microrganismo através do software JSMA

	Unknown ▼
Unknown	100%
Azospirillum amazonense	97,24011912%
Azospirillum amazonense	78,17995498%
Azospirillum amazonense	77,452228%
Escherichia coli	59,98887719%
Escherichia coli	56,80294307%
Azospirillum brasilense	51,24960208%
Escherichia coli	50,41554346%
Aeromonas hidrofila	37,67155778%
Aeromonas hidrofila	32,99855489%
Azospirillum brasilense	30,1501957%
Azospirillum brasilense	28,86045534%
Aeromonas hidrofila	28,84447075%

FONTE: O autor (2017)

4.2 CÓDIGO FONTE

Todo o código fonte do produto está hospedado no endereço eletrônico https://github.com/maltonxbr/jmsa.

A linguagem de programação utilizada, Java, mantém a organização de um código fonte através do encapsulamento de classes dentro de pacotes. Cada pacote usualmente possui classes relacionadas a uma função específica. O programa JSMA contém nove pacotes com as funções descritas na Tabela 6.

Tabela 6: Pacotes do JMSA e suas funções relacionadas

Pacote	Função e Arquivos presentes
br.ufpr.bioinfo.jmsa.analyser	Classes relacionadas com a análise ou processamento dos dados, tal como a comparação de similaridade de picos. Arquivos: • CPeaklistAnalyser.java
br.ufpr.bioinfo.jmsa.control	Classes relacionadas ao controle do funcionamento do programa, como a inicialização de variáveis globais.

	Arquivos: • CConfig.java • CControl.java • CThreadUserActionsQueue.java
br.ufpr.bioinfo.jmsa.model	Classes relacionadas às lógicas adotadas pelo programa Arquivos: • OPeak.java • OPeaklist.java
br.ufpr.bioinfo.jmsa.model.event.useraction	Classes relacionadas aos eventos disparados pelo usuário do programa que devem resultar algum processamento mais específico, tal como o carregamento dos arquivos "peaklist.xml" Arquivos: • OEvento.java • OUserActionLoadPeakFiles.java
br.ufpr.bioinfo.jmsa.resources.images	Conjunto de recursos de imagens utilizadas pelo programa, tais como os ícones de abas e botões. Arquivos: • analyser16.png • cancel16.png • cluster16.png • config32.png • folder16.png • information16.png • logo_ufpr.jpg • mainwindow32.png • ok16.png • peaklist16.png
br.ufpr.bioinfo.jmsa.resources.properties	Conjunto de propriedades que alteram alguma característica visual do programa, tal como as notas de observação Arquivos: • MESSAGES.properties
br.ufpr.bioinfo.jmsa.utils	Classes utilitárias que possuem funções de ampla utilidade pelo programa Arquivos: • CUtils.java
br.ufpr.bioinfo.jmsa.view	Classes relacionadas a interface gráfica do programa, tal como a janela de configuração Arquivos: • FAbout.java • FConfig.java • FMainWindow.java
br.ufpr.bioinfo.jmsa.view.core	Classes relacionadas a componentes

	personalizados utilizados na interface gráfica. Arquivos: • PeaklistFilesTableModel.java • PPeaklistFiles.java • PPeaklistInfo.java • PPeaklistPlot.java • PPeaklistSimilarity.java • PPeaklistTable.java • PSuperPeaklistPlot.java • SIconUtil.java

FONTE: O autor (2017)

4.3 COMPARAÇÃO DE ESPECTROS

A comparação de pares de espectros para determinar sua similaridade é uma das funções mais importantes do programa, pois ajudam um analista a escolher os espectros a comparar com maior atenção. Esta é também a primeira etapa de outras análises posteriores tais como a determinação de picos em comum para identificação de picos marcadores taxonômicos, análise de agrupamento e identificação taxonômica por busca em banco de dados. Para determinar um valor de similaridade entre dois espectros processados utilizamos alguns procedimentos, sendo eles a distância euclidiana, a função Gaussiana como distribuição normal, e a média dos valores sobre o algoritmo de Needleman-Wunsch. Estes cálculos são aplicados à comparação de cada par de picos em dois espectros, gerando uma matriz em cada etapa.

A comparação de espectros tem por entrada a lista de picos de dois espectros (arquivo “peaklist.xml”) e tem por saída uma porcentagem de zero a cem indicando a similaridade dos espectros.

Os arquivos “peaklist.xml” possuem as *tags* “mass” e “absi” para valores de razão massa/carga e intensidade, respectivamente. Considerando o que foi visto na Figura 3, os valores de razão massa/carga (m/z) tendem a se repetir enquanto as intensidades não. Dessa forma os valores de razão massa/carga representam de forma mais confiável e reprodutiva o padrão espectral dos organismos. Por esse motivo igualamos todas as intensidades dos pontos de pico e consideramos apenas os valores de ponto de m/z.

Para fins de exemplificação utilizaremos os dados exemplificados na Tabela

7, representando as listas de picos para dois espectros de massa obtidos para a bactéria *Azospirillum amazonense*, estirpe Y6. Os dois espectros foram obtidos por replicatas técnicas da amostra (duas subamostras foram geradas a partir da amostra original).

Tabela 7: Valores de intensidade de pico (ABSI) e razão massa/carga (MASS) para espectros de massa MALDI-TOF da bactéria *Azospirillum amazonense* Y6.

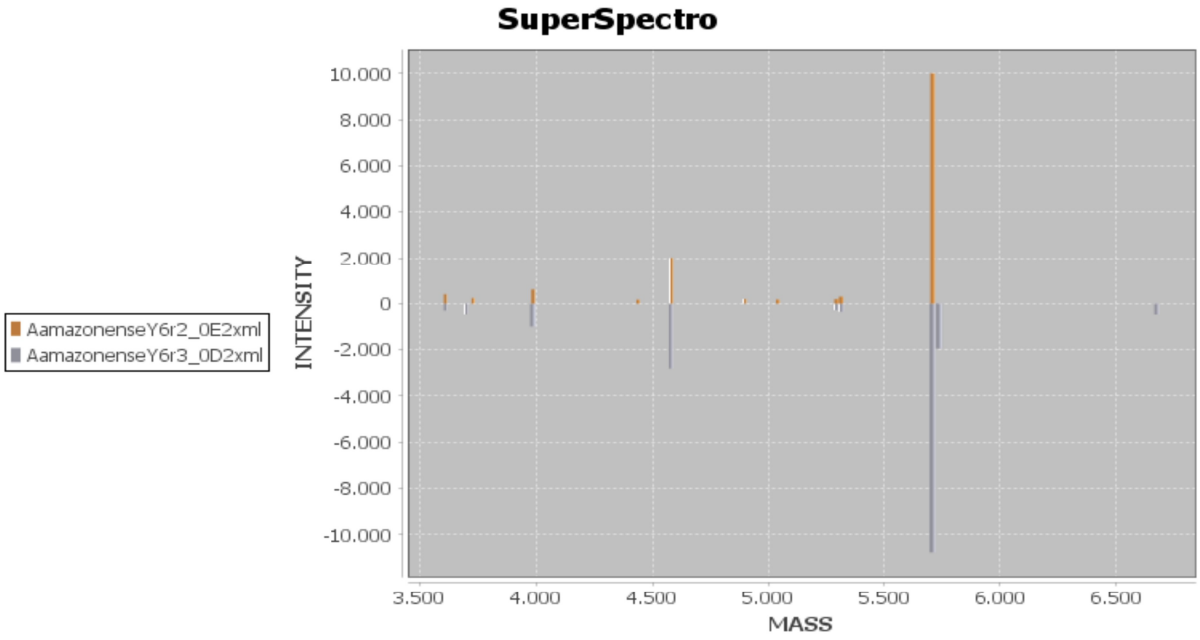
AamazonenseY6r2_0E2xml										
ABSI	412	237	615	171	2012	199	177	193	302	10001
MASS	3611	3730	3988	4438	4583	4899	5041	5294	5313	5706
AamazonenseY6r3_0D2xml										
ABSI	299	462	990	2810	282	360	10761	1940	474	-
MASS	3611	3699	3987	4583	5294	5313	5705	5734	6673	-

FONTE: O autor (2017)

LEGENDA: ABSI - Valores de intensidade dos picos
MASS - Valores da razão m/z dos picos.

Esses mesmos dados, apresentados na Tabela 7, podem ser representados no programa JMSA através do gráfico mostrado na Figura 19.

Figura 19: Amostra de dados espelhada graficamente.



FONTE: O autor (2017)

4.3.1 Distância euclidiana

Os valores de distância euclidiana são calculados para cada par de picos em dois espectros. A Tabela 8 mostra a matriz de distâncias para a comparação entre os espectros usados como exemplo na Tabela 7 e Figura 19. Para efeito de otimização de código, evitando processamento computacional desnecessário, o algoritmo considera apenas o valor quadrático da distância euclidiana, uma vez que, na etapa seguinte, o quadrado das distâncias euclidianas são considerados para o cálculo da distribuição normal (ver abaixo).

Tabela 8: Resultado do cálculo de distância euclidiana sobre a matriz de picos para os espectros de *Azospirillum amazonense,* mostrados na Tabela 7

		AamazonenseY6r3_0D2xml								
	MASS	**3611**	**3699**	**3987**	**4583**	**5295**	**5313**	**5705**	**5734**	**6673**
AamazonenseY6r2_0E2xml	**3611**	0	88	376	972	1683	1702	2094	2123	3062
	3730	119	31	257	853	1564	1583	1975	2004	2943
	3988	377	289	1	595	1306	1325	1717	1746	2685
	4438	827	739	451	145	856	875	1267	1296	2235
	4583	972	884	596	0	711	730	1122	1151	2090
	4899	1288	1200	912	316	395	414	806	835	1774
	5041	1430	1342	1054	458	253	272	664	693	1632
	5294	1683	1595	1307	711	0	19	411	440	1379
	5313	1702	1614	1326	730	19	0	392	421	1360
	5706	2095	2007	1719	1123	412	393	1	28	967

FONTE: O autor (2017)

4.3.2 Distribuição normal

Ao obtermos a distância euclidiana dos pontos de massa podemos observar um valor quantitativo linear que define os pontos de picos que mais se assemelham (mais próximos ou com distâncias menores), entretanto ainda é necessário estabelecer uma curva de distribuição para implicar maior peso sobre valores que mais diferem entre si. Para isso sujeitamos a matriz representada na Tabela 8 à função de distribuição normal, sendo que o desvio padrão utilizado serve de filtro para determinar a porcentagem de semelhança entre pontos de picos.

Como exemplo, levemos em consideração a comparação entre os picos de valores m/z 3611 (AamazonenseY6r2_0E2xml) e 3699 (AamazonenseY6r3_0D2xml).

Temos que o quadrado da distância euclidiana desses pontos é:

$$d(3611,3699)^2=\left(\sqrt{(x_i-\ y_i)^2}\right)^2=(3611-\ 3699)^2=7744$$

Uma vez que esse já é o valor quadrático da distância, ao aplicarmos na função Gaussiana temos que:

- "a" é o valor com mais presença, no caso 1 é o valor máximo de presença
- "x" é o valor a ser comparado (3699 em relação à 3611)
- "b" é o valor sobre a qual a perspectiva está (3611), portanto considerado a mediana, a média e o centro da curvatura.
- "c" é o desvio padrão estabelecido no programa de forma estática de valor 256

$$f(3699,3611,256)=1*\exp\left(\frac{-(3699-\ 3611)^2}{2*(256^2)}\right)$$

$$f(3699,3611,256)=\exp\left(\frac{-\ 7744}{2*(256^2)}\right)=0{,}9426294409$$

Esse cálculo é realizado na comparação de cada pico para preencher a Tabela 9.

Tabela 9: Resultado da função Gaussiana sobre a matriz de distância euclidiana, calculada na Tabela 8, dos pares de picos para os espectros de *Azospirillum amazonense*.

		AamazonenseY6r3_0D2xml								
	MASS	**3611**	**3699**	**3987**	**4583**	**5295**	**5313**	**5705**	**5734**	**6673**
AamazonenseY6r2_0E2xml	**3611**	1.00000	0.94263	0.34007	0.00074	0.00000	0.00000	0.00000	0.00000	0.00000
	3730	0.89759	0.99269	0.60416	0.00388	0.00000	0.00000	0.00000	0.00000	0.00000
	3988	0.33812	0.52876	0.99999	0.06714	0.00000	0.00000	0.00000	0.00000	0.00000
	4438	0.00542	0.01551	0.21186	0.85180	0.00373	0.00291	0.00000	0.00000	0.00000
	4583	0.00074	0.00257	0.06653	1.00000	0.02114	0.01715	0.00007	0.00004	0.00000
	4899	0.00000	0.00002	0.00175	0.46681	0.30411	0.27046	0.00704	0.00490	0.00000
	5041	0.00000	0.00000	0.00021	0.20182	0.61364	0.56867	0.03460	0.02563	0.00000
	5294	0.00000	0.00000	0.00000	0.02114	1.00000	0.99725	0.27561	0.22831	0.00000
	5313	0.00000	0.00000	0.00000	0.01715	0.99725	1.00000	0.30963	0.25866	0.00000
	5706	0.00000	0.00000	0.00000	0.00007	0.27389	0.30779	0.99999	0.99404	0.00080

FONTE: O autor (2017)

4.3.3 Cálculo da similaridade entre dois espectros

Para transformar a tabela de valores de similaridade num único valor que representa o conjunto da comparação (similaridade entre os dois espectros) era necessário determinar os picos que seriam comuns a ambos os espectros, e assim evitar valores duplicados, para assim extrair um valor médio.

Para esse fim utilizamos somente a etapa final do algoritmo Needleman-Wunsch (*traceback*), utilizando a própria matriz gerada pela gaussiana como matriz de similaridade e nenhuma penalidade para inserções e deleções. Desta forma, para determinar a similaridade média entre os espectros usados no exemplo, o "caminho" contendo os picos correspondentes nos dois espectros é traçado a partir do valor na célula da última coluna e última linha à direita na tabela até o valor na célula da primeira coluna e primeira linha à esquerda na tabela, recuperando sempre as células com os maiores valores no traçado (destacados em verde na Tabela 10). O *tracebacking* no exemplo gera um valor de similaridade para os espectros de 0,82436 ou 82,436%.

É importante notar que o algoritmo Needleman-Wunsch foi originalmente usado para o alinhamento global entre duas sequências de nucleotídeos ou aminoácidos e, neste caso, considera a inserção de *gaps* no resultado final, sendo que cada inserção de *gap* penaliza o valor do *score* total. Aqui o objetivo do *traceback* não é gerar um alinhamento entre dois espectros, uma vez que os valores de m/z são fixos em cada espectro, não podendo ser deslocados com a inserção de *gaps* para otimização de um alinhamento, mas sim determinar quais são os picos equivalentes (de acordo com as distâncias euclidianas) presentes nos dois espectros, tolerando um deslocamento dentro da distribuição normal. Desta forma, as regiões que conteriam *gaps* indicam picos presentes em um dos espectros e ausentes no outro. Por exemplo, na Tabela 10, a coluna para o pico 4583 de AamazonenseY6r3_0D2xml indica, no *traceback*, possuir correspondência com os picos 4438, 4583 e 4899 de AamazonenseY6r2_0E2xml. Nota-se que a maior similaridade ocorre entre os picos 4583 em ambos os espectros, mas valores de similaridade são gerados para os demais picos em função da distância euclidiana. No exemplo dado, os dois espectros possuem o pico 4583, sendo a similaridade neste ponto máxima, igual a 1,0. Entretanto os picos 4438 e 4899 no espectro

AamazonenseY6r2_0E2xml estão ausentes no espectro AamazonenseY6r2_0D2xml e a similaridade entre eles é calculada em relação ao pico 4583, o mais próximo. No cálculo da média de similaridades entre picos, os valores para picos ausentes em um dos espectros constituem um forma de "penalização", pois serão calculados em relação a picos em posição m/z não equivalentes.

Tabela 10: Etapa de *tracebacking* dos valores de picos que serão incluídos na média de similaridade para a comparação dos pares de picos para os espectros de *Azospirillum amazonense* mostrados na Tabela 7.

		AamazonenseY6r3_0D2xml								
	MASS	**3611**	**3699**	**3987**	**4583**	**5295**	**5313**	**5705**	**5734**	**6673**
AamazonenseY6r2_0E2xml	**3611**	1.00000	0.94263	0.34007	0.00074	0.00000	0.00000	0.00000	0.00000	0.00000
	3730	0.89759	0.99269	0.60416	0.00388	0.00000	0.00000	0.00000	0.00000	0.00000
	3988	0.33812	0.52876	0.99999	0.06714	0.00000	0.00000	0.00000	0.00000	0.00000
	4438	0.00542	0.01551	0.21186	0.85180	0.00373	0.00291	0.00000	0.00000	0.00000
	4583	0.00074	0.00257	0.06653	1.00000	0.02114	0.01715	0.00007	0.00004	0.00000
	4899	0.00000	0.00002	0.00175	0.46681	0.30411	0.27046	0.00704	0.00490	0.00000
	5041	0.00000	0.00000	0.00021	0.20182	0.61364	0.56867	0.03460	0.02563	0.00000
	5294	0.00000	0.00000	0.00000	0.02114	1.00000	0.99725	0.27561	0.22831	0.00000
	5313	0.00000	0.00000	0.00000	0.01715	0.99725	1.00000	0.30963	0.25866	0.00000
	5706	0.00000	0.00000	0.00000	0.00007	0.27389	0.30779	0.99999	0.99404	0.00080

FONTE: O autor (2017)

4.4 PERFORMANCE

O programa foi testado para performance e os resultados são mostrados na Tabela 11. Foi verificado que o aumento no tempo gasto é cerca de 47 vezes para importação de 1.000 arquivos, quando comparado à importação de 10 arquivos. Este aumento foi bem maior que o aumento verificado para consumo de memória, em torno de 3,5 vezes. O aumento verificado para uso da CPU também foi bastante significativo, atingindo 50 vezes (em porcentagem). Entretanto, os valores absolutos para importação de 1.000 arquivos não foram críticos, sendo menor que 360 KB de memória RAM e 30% de CPU.

Tabela 11: Teste de performance do JMSA.

Arquivos carregados	10	100	1000
Tempo gasto (ms)	161	809	7640

Uso de memória (KB)	101,480	137,924	358,832
Uso médio de CPU (%)	0,6%	9,0%	30,0%

FONTE: O autor (2017)

4.5 PROBLEMAS CONHECIDOS

Para esta versão do programa, algumas limitações na sua funcionalidade para o usuário final são listadas abaixo:

Incapacidade de carregar arquivos de pastas separadas individualmente.

Selecionar aba de Informações quando há múltiplos espectros selecionados permite alterar apenas o primeiro selecionado.

Descrição em arquivo de informações de espectros não carrega corretamente

Não há um indicativo de sucesso ao salvar as informações de um espectro

Não há modo de configurar técnicas de comparação ou parâmetros como a tolerância de distância entre picos, medido em Daltons.

5 CONCLUSÃO

Neste trabalho foi desenvolvido o programa JMSA para gerenciamento, visualização e análise de espectros de massa MALDI-TOF para identificação de microrganismos. As principais características do programa, são:

1. Multiplataforma;
2. Código aberto;
3. Capacidade para importação, manipulação e análise de centenas a milhares de espectros de massa ao mesmo tempo;
4. Interface gráfica fácil e intuitiva.

A partir do programa desenvolvido, outras implementações podem ser feitas em versões futuras para melhorar a usabilidade e utilidade do programa para manipulação, visualização e análise dos espectros de massa. A principal implementação sugerida é a criação de um banco de dados interno configurável para cada projeto iniciado e manter espectros que componham um banco de dados de organismos referência para a identificação a partir de espectros de organismos desconhecidos. Neste banco de dados, podemos ter a opção de incluir superespectros dos organismos representados, de forma a garantir uma maior confiança no resultado das comparações.

Também é necessário implementar outras ferramentas analíticas como a criação de dendrogramas, a aplicação de outros métodos para o cálculo de similaridade entre picos e espectros e a implementação de métodos para detecção de picos a partir dos dados brutos, eliminando a dependência do pré-processamento no programa proprietário FlexAnalysis. Isto também ampliaria o número de sinais que poderiam ser usados para a caracterização dos organismos, com ganho na precisão e confiabilidade das análises.

Além disso, a utilização de formatos de arquivo abertos, como o mzXML poderia permitir a utilização do programa de forma mais ampla e independente da plataforma de espectrometria de massa utilizada nas análises das amostras.

REFERÊNCIAS

Alm, R. et al. Detection and Identification of Protein Isoforms Using Cluster Analysis of MALDI−MS Mass Spectra. Journal of Proteome Research, v. 5, n. 4, p. 785-792, 2006.

Aston, F. LXXIV. A positive ray spectrograph. Philosophical Magazine Series 6, v. 38, n. 228, p. 707-714, 1919.

BEYNON, J. H. (John Herbert). Mass spectrometry and its applications to organic chemistry. Amsterdam: Elsevier, 1960. xiii, 640 p., il..

BIEMANN, K. (Klaus). Mass spectrometry: organic chemical applications. New York: McGraw-Hill, c1962. 370 p., il. (McGraw-Hill series in advanced chemistry).

Com Ciência - Bioinformática. Disponível em: <http://www.comciencia.br/reportagens/bioinformatica/bio03.shtml>. Acesso em: 17 ago. 2016.

Communiqué - Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry for the Identification of Bacterial and Yeast Isolates - Mayo Medical Laboratories. Disponível em: <http://www.mayomedicallaboratories.com/articles/communique/2013/01-maldi-tof-mass-spectrometry/>. Acesso em: 17 ago. 2016.

Croxatto, A.; Prod'hom, G.; Greub, G. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS Microbiology Reviews, v. 36, n. 2, p. 380-407, 2012.

Cruz, C. D., Carneiro, P. C. S. (2003) Modelos biométricos aplicados ao melhoramento genético. Viçosa: UFV, vol. 2, 585p.

Demirev, P.Fenselau, C. Mass Spectrometry for Rapid Characterization of Microorganisms. Annual Review of Analytical Chemistry, v. 1, n. 1, p. 71-93, 2008.

Dempster, A. A new Method of Positive Ray Analysis. Physical Review, v. 11, n. 4, p. 316-325, 1918.

du Bois, H. Ueber magnetische Schirmwirkung. Annalen der Physik, v. 301, n. 5, p. 1-37, 1898.

Falconer, I. J J Thomson and the discovery of the electron. Physics Education, v. 32, n. 4, p. 226-231, 1997.

Grayson, M. Measuring mass. Traducao . 1. ed. Philadelphia: Chemical Heritage Press, 2002.

Karas, M.; Bachmann, D.; Hillenkamp, F. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. Analytical Chemistry, v. 57, n. 14, p. 2935-2939, 1985.

López-Fernández, H. et al. Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery. BMC Bioinformatics, v. 16, n. 1, 2015.

MALLET, A. I. Dictionary of mass spectrometry. Chichester; Hoboken, NJ: Wiley, 2009. vii, 174 p., il. ISBN 0470027614.

MCDOWELL, Charles A. Mass spectrometry. New York: McGraw-Hill Book, c1963. x, 639p., il. (McGraw-Hill series in advanced chemistry).). Inclui bibliografia e indice.

MCLAFFERTY, F. W. Mass spectral correlations. Washington, D.C.: ACS, 1963. 117 p. (Advances in chemistry series, 40).

Needleman, S.Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, v. 48, n. 3, p. 443-453, 1970.

Parsons, L. Reading Mass Spectra. Disponível em: <http://parsonsfamily.boldlygoingnowhere.org/~lparsons/upstate/massspec/ch4/>. Acesso em: 17 ago. 2016.

Rahi, P.; Prakash, O.; Shouche, Y. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass-Spectrometry (MALDI-TOF MS) Based Microbial Identifications: Challenges and Scopes for Microbial Ecologists. Frontiers in Microbiology, v. 7, 2016.

Santos, I.; Hildenbrand, Z.; Schug, K. Applications of MALDI-TOF MS in environmental microbiology. The Analyst, v. 141, n. 10, p. 2827-2837, 2016.

STETS, M.; PINTO, A.; HUERGO, L. et al. Rapid identification of bacterial isolates from wheat roots by high resolution whole cell MALDI-TOF MS analysis. Journal of Biotechnology, v. 165, n. 3-4, p. 167-174, 2013.

Tan, K. et al. Prospective Evaluation of a Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry System in a Hospital Clinical Microbiology Laboratory for Identification of Bacteria and Yeasts: a Bench-by-Bench Study for Assessing the Impact on Time to Identification and Cost-Effectiveness. Journal of Clinical Microbiology, v. 50, n. 10, p. 3301-3308, 2012.

Tanaka, K. et al. Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry, v. 2, n. 8, p. 151-153, 1988.

The Nobel Prize in Chemistry 1922. Disponível em: <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1922/>. Acesso em: 23

maio. 2017.

The Nobel Prize in Chemistry 2002. Disponível em: <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2002/>. Acesso em: 23 maio. 2017.

The Nobel Prize in Physics 1906. Disponível em: <http://www.nobelprize.org/nobel_prizes/physics/laureates/1906/>. Acesso em: 23 maio. 2017.

Thomson, J. XIX.Further experiments on positive rays. Philosophical Magazine Series 6, v. 24, n. 140, p. 209-253, 1912.

Thomson, J. XL.Cathode Rays. Philosophical Magazine Series 5, v. 44, n. 269, p. 293-316, 1897.

Ward, J. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, v. 58, n. 301, p. 236-244, 1963.

WATSON, J. Throck. Introduction to mass spectrometry: instrumentation, applications and strategies for data interpretation. 4th ed. Chichester, England; Hoboken, NJ: John Wiley & Sons, c2007. xxiv, 818 p., ill., 26 cm. ISBN 0470516348 (cloth).

Welham, K.; Domin, M.; Scannell, D.; Cohen, E.; Ashton, D. The characterization of micro-organisms by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry, v. 12, n. 4, p. 176-180, 1998.

Williams, T. et al. Experimental factors affecting the quality and reproducibility of MALDI TOF mass spectra obtained from whole bacteria cells. Journal of the American Society for Mass Spectrometry, v. 14, n. 4, p. 342-351, 2003.

Wunschel, S. et al. Bacterial analysis by MALDI-TOF mass spectrometry: An inter-laboratory comparison. Journal of the American Society for Mass Spectrometry, v. 16, n. 4, p. 456-462, 2005.

Yamashita, M.; Fenn, J. Electrospray ion source. Another variation on the free-jet theme. The Journal of Physical Chemistry, v. 88, n. 20, p. 4451-4459, 1984.