República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(11) (21) **PI 0506117-2 A**

(22) Data de Depósito: 14/10/2005
(43) Data de Publicação: **03/07/2007**
(RPI 1904)

(51) Int. Cl.:
G01N 33/50 (2007.01)
G01N 33/574 (2007.01)

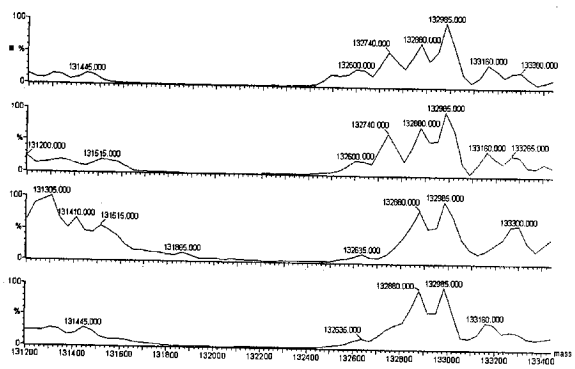
(54) Título: **MÉTODO DE DIAGNÓSTICO BASEADO EM PADRÕES PROTEÔMICOS E/OU GENÔMICOS POR VETORES DE SUPORTE APLICADO A ESPECTOMETRIA DE MASSA**

(71) Depositante(s): Fundação Oswaldo Cruz (BR/RJ)

(72) Inventor(es): Wim Maurits Sylvain Degrave, Paulo Costa Carvalho, Maria da Glória da Costa Carvalho, Gilberto Barbosa Domont, Raul Fonseca Neto, Sergio Lilla

(74) Procurador: Bhering, Almeida & Associados

(57) Resumo: Método de diagnóstico baseado em padrões proteômicos e/ou genômicos por vetores de suporte aplicado a espectometria de massa A presente invenção refere-se a um método de diagnóstico médico baseado em padrões proteômicos e/ou genômicos, por meio de dados obtidos em espectros gerados pela espectometria de massa no qual é possível classificar os pacientes quanto ao estágio de infectabilidade. Adicionalmente, a presente invenção também se refere a dois novos biomarcadores para o diagnóstico médico da Doença de Hodgkin. Baseado nas análises de SVM localiza-se as janelas de interesse e posteriormente recorre-se ao espectro de massa para que ocorra a localização dos biomarcadores, de modo que a identificação dos ditos biomarcadores ocorra por meio de um gel 2D ou por meio da espectrometria de massa.



PI0508117

Método de diagnóstico baseado em padrões proteômicos e/ou genômicos por vetores de suporte aplicado a espectometria de massa.

Campo da Invenção

A presente invenção refere-se a um método de diagnóstico médico baseado em padrões proteômicos e/ou genômicos, por meio de dados obtidos em espectros gerados pela espectometria de massa no qual é possível classificar os pacientes quanto ao estágio de infectabilidade. Adicionalmente, a presente invenção também se refere a dois novos biomarcadores para o diagnóstico médico da Doença de Hodgkin.

Fundamentos da Invenção:

Durante os últimos 40 anos, a possibilidade de detecção do câncer num estágio inicial por meio de biomarcadores permitiu uma imensa transformação no campo de diagnóstico médico, além de permitir uma quantificação na progressão da resposta do individuo ao tratamento.

A busca por um biomarcador ocorre por meio de análises do proteoma do paciente, entretanto, para distinguir uma única patologia, as análises de um, ou de alguns biomarcadores tem sido sem sucesso até os dias de hoje. Como exemplo de biomarcador pode-se citar o biomarcador PSA. Dito biomarcador é um antígeno especifico utilizado no diagnóstico de câncer de próstata.

Proteoma é definido como proteínas expressas pelo genoma apresentado, o qual pode variar intensamente com o tempo, com a doença ou com o tipo de droga utilizada

durante o tratamento e as diversas análises proteômicas que compõem o quadro de técnicas de estudo do proteoma.

A maior parte das análises do proteoma, difundida para biomarcadores ocorre por meio da técnica da análise bidimensional de um gel de eletroforese (2DE) de alta resolução oriundo das amostras provenientes do material coletado de pacientes infectados e de amostras provenientes do material coletado de pacientes saudáveis. Entretanto, embora esta técnica em apreço, juntamente com técnicas complementares, tenha contribuído para o desenvolvimento do campo genômico, ainda existem muitas limitações no estado da arte de modo a prever a capacidade codante de um genoma, e do proteômico, a fim de evidenciar a presença e a localização protéica na célula, identificar marcadores de doenças e alvos de medicamentos. A rotina técnica 2DE não é adequada para atuar como um diagnóstico médico, uma vez que a métodologia de análise é um processo demorado, laborioso, restrito somente à observação de proteínas em uma determinada faixa de pH a qual varia de 3.5-11.5, desnaturadas e com peso molecular variando de 7-200 kDa. Além do mais, mesmo para rastrear os biomarcadores, os géis 2D teriam que ser aplicados a um número grande de amostras, de modo a gerar altíssimos custos para este tipo de pesquisa. Uma das vantagens do gel 2D, é que dependendo das condições de fracionamento e da espécie biológica investigada, esta técnica permite a separação de diversas proteínas por espectrometria de massa. Uma vez localizada qual proteína está diferencialmente expressa, esta pode ser cortada do gel de eletroforese, digerida com tripsina de

modo que o resultado desta digestão triptica resulte em um padrão de massas, ou impressão digital molecular desta, a fim de permitir a sua identificação por meio de um espectrometro de massas preferencialmente, por meio da técnica de Ionização por desorção a laser auxiliada por matriz, do inglês, *Matrix Assisted Laser Desorption Ionization* (MALDI) devido a sua simplicidade de uso. A intensidade de certos picos diferencialmente expressos, correspondente a proteínas diferencialmente expressas podem caracterizar potenciais marcadores, os quais podem ser utilizados para o diagnóstico antecipado de algumas infectabilidades.

Na literatura, são descritos diversos métodos para caracterização de novos biomarcadores. Alguns dos métodos descritos têm sido utilizados para distinguir amostras de câncer de amostras controles "saudáveis" por meio direto dos dados das análises do proteoma.

Recentemente, têm-se utilizado a técnica SELDI - TOF, a qual por meio de dados obtidos no espectro de massa, correlacionados com um algoritmo de computador permite a identificação de proteínas chaves para o diagnóstico de pacientes com câncer no ovário. Uma outra técnica baseia-se nas análises de máxima separabilidade, as quais são utilizadas para diagnosticar pacientes com câncer no tórax.

Entretanto, a busca por um biomarcador para distinguir uma única patologia tem sido sem sucesso até os dias de hoje. A depleção das proteínas pode resultar na perda de biomarcadores em potencial ou na troca dos soros conjugados.

A doença de Hodgkin (DH) foi uma espécie de laboratório em que se desenvolveram importantes inovações conceituais que fazem parte da abordagem diagnóstica e terapêutica atual em todo o campo da oncologia médica. Desta forma, ela foi utilizada como modelo de estudo na presente concretização.

A doença de Hodgkin é caracterizada pela presença de linfomas nas células do organismo de um paciente.

Para o diagnóstico clínico da Doença de Hodgkin é necessário que o individuo se submeta a diversos tipos de exames, de modo que ditos exames permitam determinar o subtipo específico entre outras informações úteis para se decidir o tratamento mais adequado para o paciente.

Um método de diagnóstico proposto nas técnicas relacionadas é descrito na patente US 6.835.927. Na dita patente é descrita a análise de picos expressos diferentemente em um espectro de massa por meio da técnica do PCA (Principal Component Analysis). Dita técnica é baseada em uma matriz de correlação, de modo a não ser uma solução ótima para casos onde exista uma grande heterogeneidade em análise, como por exemplo, no caso do soro humano.

Para localizar os "marcadores" são sugeridos diversas abordagens, como por exemplo, o método dos mínimos quadrados e análise do principal componente redutor dos diversos picos para um reduzido grupo de picos. Nestes espectros de massa podem-se comparar as intensidades dos picos de maneira a se obter um diagnóstico para o paciente.

O método dos mínimos quadrados, assim como o de redes neurais, e os classificadores clássicos da estatística baseiam-se na minimização de uma função de erro. É provado que para fazer o erro desta função tender a zero, o número de casos estudados deve tender ao infinito.

Entretanto, o método descrito na patente US 6.835.927, não deixa evidente a maneira utilizada para classificar um paciente doente ou um paciente "saudável" quando ocorre de os picos obtidos nos espectros de massa não refletirem o padrão esperado. Uma outra desvantagem do método descrito é a limitada abrangência do conjunto de picos. Devido a grande complexidade e inúmeras variações nos fluidos biológicos, especialmente no soro humano, um diagnóstico baseado em um número limitado de picos, desde que não propriamente selecionados, não reflete de maneira correta o diagnóstico do paciente, assim como a possibilidade de estimar a certeza do diagnóstico.

Problemas como estes aqui apresentados envolvendo uma alta dimensão necessitam de um número exponencial de parâmetros em relação à sua dimensionalidade. Como por exemplo; em uma função discriminante quadrática possui uma quantidade de parâmetros ao quadrado da dimensão dos dados quando em uma máquina de classificação linear. Diversos algoritmos de aprendizagem de máquinas dependem do conhecimento a priori (dos métodos estatísticos Bayes) de uma distribuição de probabilidade, tornando assim um fator limitante no método. Por outro lado, os modelos de regressão baseados no critério MAP máxima probabilidade a posteriori se baseiam no principio de minimização de risco

empírico, ou seja, procuram minimizar uma função de erro que é uma aproximação discreta do erro verdadeira.

Entretanto, o método descrito na patente US 6.835.927, não deixa evidente a maneira utilizada para classificar um paciente doente ou um paciente saudável quando ocorre de os picos obtidos nos espectros de massa não refletirem o padrão esperado. Uma outra desvantagem do método descrito é a limitada abrangência do conjunto de picos. Devido a grande complexidade e inúmeras variações nos fluidos biológicos, especialmente no soro humano, um diagnóstico baseado em um número limitado de picos não reflete de maneira correta o diagnóstico do paciente.

Um outro método encontrado nas técnicas relacionadas encontra-se descrito no documento de patente US 6.134.344. Este documento descreve um método para melhorar a eficiência da análise, de modo a utilizar um conjunto reduzido de vetores e de maneira a mapear os ditos vetores em um espaço de maior dimensão durante a fase de treinamento.

O método de kernels apesar de ser uma solução eficiente para resolução de diversos problemas matemáticos como o XOR, suas aplicações em ciências da saúde ainda não são profundamente compreendidas e assim, pode-se facilmente tomar "decisões viciadas" a partir de um conjunto de dados. O conjunto de vetores reduzidos não é composto por vetores de suporte, nem vetores utilizados durante o teste.

O método de diagnóstico descrito na presente invenção permite a realização de uma redução do espaço de características, de modo a utilizar uma seleção de

características por meio do método de máxima divergência (SVM). Dito método permite classificar dados não linearmente separáveis no espaço de características.

O método SVM é uma classe de algoritmos, caracterizada pela utilização de kernel, ausência de mínimo local, solução esparsa e caracterizada por vetores de suporte baseado na teoria de minimização de risco estrutural. Em problemas complexos, modelos que possuem uma alta capacidade de "ajuste" ao conjunto de dados de treinamento poderiam ter "aprendizado viciado" denominado de *overfitting,* e perderiam o poder de generalização.

O método SVM supera as metodologias previamente existentes em diversas áreas consagradas como: reconhecimento de imagens, categorização de textos, reconhecimento de caligrafia, ou de padrões sonoros e de problemas os quais dificilmente podem ser modelados matemáticamente. Recentemente, com a aplicação do método SVM na Bioinformática, foi possível analisar dados de micro-arranjo de câncer de modo a permitir a caracterização do estágio de evolução da doença, desenho de novos fármacos, classificação das proteínas quanto à função, previsão dos seus formatos, estrutura e localização subcelular, interações proteína-proteína e reconhecimento das proteínas transmembranares entre outras.

Como vantagem, a abordagem do método da presente invenção permite avaliar as modificações pós-traducionais, nas quais o método PCA apresenta falha ou produz resultados não satisfatórios. Baseado no método SVM localiza-se as janelas de interesse e posteriormente recorre-se ao

espectro de massa para que ocorra a localização dos biomarcadores, de modo que a identificação dos ditos biomarcadores ocorra por meio de um gel 2D ou por meio da espectrometria de massa.

SUMÁRIO DA INVENÇÃO

A presente invenção apresenta um método de diagnóstico médico baseado em padrões proteômicos e/ou genômicos, por meio de dados obtidos em espectros gerados pela espectometria de massa no qual é possível classificar os pacientes quanto ao estágio de infectabilidade.

Dito método de diagnóstico permite a realização de uma redução do espaço, de modo a utilizar uma seleção de características por meio da análise de máxima divergência (SVM). A análise SVM "navega" sobre o pool de espectros de massa e com a utilização do método de "cross validation", do tipo "leave-one-out", obtém os trechos do espectro de maior relevância e variabilidade, a fim de reduzir assim o sinal/ruído no processo decisório de diagnóstico, o qual baseia-se na classificação dos soros em positivos, negativos, falsos positivos e falso negativo.

Adicionalmente, a presente invenção também se refere à descoberta de dois novos biomarcadores para o diagnóstico médico da Doença de Hodgkin. Baseado nas análises de SVM localiza-se as janelas de interesse e posteriormente recorre-se ao espectro de massa para que ocorra a localização dos biomarcadores, de modo que a identificação dos ditos biomarcadores ocorra por meio de um gel 2D ou por meio da espectrometria de massa.

Assim, um primeiro objetivo da presente invenção é proporcionar um método de diagnóstico médico, o qual por meio de dados obtidos em espectros gerados pela espectometria de massa é possível classificar os pacientes quanto ao estágio de infectabilidade.

Um outro objetivo da presente invenção se refere a dois novos biomarcadores para o diagnóstico médico da Doença de Hodgkin.

BREVE DESCRIÇÃO DAS FIGURAS

A Figura 1 exemplifica um plano sendo utilizado como superfície decisória entre duas classes de pontos.

A Figura 2 mostras os resultados de MDA para uma abertura da janela de navegação aproximada de 2240 e 4480 Daltons.

A Figura 3 mostra o espectro de massa de picos expressos em aproximadamente 1327400Da para amostras de sangue de pacientes controles.

A Figura 4 mostra as análises de MDA para a janela de estudo de aproximadamente 20 m/z e 10 m/z. Esta região do espectro indica o sítio indicativo de potenciais biomarcadores para diagnóstico clínico.

A Figura 5 mostra o espectro de massa para uma região do espectro, no qual observa-se a presença de envelopes isotópicos expressos diferentemente em aproximadamente 980 e 994 m/z em amostras de sangue de pacientes controles.

DESCRIÇÃO DETALHADA DA INVENÇÃO

A presente invenção mostra a solução para os problemas existentes no estado da arte na forma de um gráfico, no

qual os biomarcadores e regiões de importância do espectro massa podem ser facilmente visualizados.

O método da presente invenção descreve a busca por biomarcadores, entretanto, dita busca não se baseia em comparar picos do espectro de massa exata, mas a comparação de uma reduzida janela de estudo na qual analisa-se um perfil do espectro por vez.

Baseado na análise de SVM localiza-se as janelas de interesse e posteriormente recorre-se ao espectro de massa para que ocorra a localização dos biomarcadores, de modo que a identificação dos ditos biomarcadores ocorra por meio de um gel 2D ou por meio da espectrometria de massa.

Na espectrometria de massa busca-se não apenas um pico diferencialmente apenas expresso, o qual pode ser um mero spike ou contaminante, mas busca-se preferencialmente um pequeno perfil proteômico expresso por meio dos picos do espectro, que podem ser verificados através de seu envelope isotópico. Este fato aumenta a certeza de se obter um biomarcador protéico, de modo a se prescrever diagnósticos médicos baseados no perfil proteômico, caracterizar e elaborar kits para imunodiagnósticos.

Para superar os limites do estado da arte, tendo como pilar principal a análise de SVM, ou o classificador de máxima margem, a invenção é capaz de lidar com a esparsidade, escassez do conjunto de treinamento e assumir o não conhecimento a priori, da quantidade de parâmetros necessários ao modelo.

O método da presente invenção baseia-se no principio da minimização do risco estrutural, um novo princípio de

indução derivada da teoria de aprendizagem estatística introduzida por Vapnik e Chervonenkis sendo uma evolução da teoria anterior da minimização do risco empírico (ERM).

Assim, a presente invenção apresenta um método para evitar a perda de biomarcadores em potencial, através do uso da técnica de espectrometria de massa, a qual faz uso da técnica de ionização por spray de elétrons, de modo a permitir uma ionização na fase liquida para a fase gasosa de uma maior quantidade de proteinas séricas, de modo a permitir a análise por espectrometria de massa.

Adiante, conforme será demonstrado, será aplicada a metodologia de máquinas de vetores de suporte para classificar uma amostra como de doente ou saudável, baseado em todo o perfil proteômico obtido no espectrômetro de massas.

A análise SVM utilizada para reduzir a sinal / ruído ruído também poderá ser aplicada (máxima divergência por vetores de suporte) para otimizar o processo decisório. Como continuação do "filtro" de máxima divergência, os dados selecionados como mais importantes também podem contribuir para um aprimoramento da construção de uma superfície controle elíptica que tenta separar novamente pessoas saudáveis de pessoas doentes. A vantagem desta nova superfície é que ela é otimizada para a resolução de problemas com mais de uma classe (Controle vs cancer de prostata, vs pulmão vs cabeça e pescoço).

A invenção será agora descrita com base em exemplos, os quais não devem ser considerados limitativos da mesma.

Exemplo 1: Coletas das amostras de sangue

Foram coletadas 30 amostras de sangue de pacientes saudáveis 30 amostras de sangue de pacientes infectados com a Doença de Hodgkin imediatamente após o diagnóstico médico e antes de iniciado o tratamento.

A classificação histológica das amostras de sangue foi confirmada por hematologistas, de acordo com o critério da OMS (Organização Mundial da Saúde/World Health Organization-WHO).

A presença do vírus Epstein-Barr (EBV) nas células tumorais foi avaliada por meio da expressão imunohistoquímica da proteína LMP - 1 (proteína latente de membrana) com o uso do coquetel monoclonal de anticorpo CS1-4.

A avaliação dos pacientes incluiu o histórico completo do paciente, exame físico, contagens diversas e completas das amostras de sangue, arquivos bioquímicos, sorologia para HIV, radiografia do tórax, tomografia computadorizada do tórax, abdômen e biopsia da medula óssea.

O soro extraído das amostras de sangue dos pacientes foi armazenado em alíquotas a uma temperatura aproximada de -80 °C. A demografia tumoral, o desenvolvimento do tumor, e as demais informações patológicas dos pacientes, foram armazenadas em um banco de dados de computador.

Exemplo 2: Análise do proteoma

Antes de serem iniciadas as análises do proteoma do paciente, as alíquotas dos soros obtidos a partir de cada uma das amostras de sangue coletadas dos dois grupos de pacientes, foram descongeladas a uma temperatura

preferencial de aproximadamente 25°C e levada a agitação por meio de um vórtex.

A cada uma das alíquotas de soro foi realizada uma diluição de aproximadamente 1:3, preferencialmente em água Milli-Q e as ditas alíquotas de soro foram dessalinizadas preferencialmente em membrana Millipore™ do tipo zip-tip provida de uma membrana C4 em sua extremidade (290 Concord Road, Billerica, MA 01821, USA) de acordo com o manual do fabricante.

A solução da alíquota final, continha cerca de 2µL de soro, ou seja, a mistura de proteínas sofreu uma nova diluição, de modo a atingir um volume aproximado de 10µL por meio da adição da amostra de sangue na solução preparada.

Exemplo 3: Obtenção dos espectros

Todos os espectros de massa foram obtidos por meio de um espectrômetro de massa o qual combina dois analisadores: uma fonte ionizadora do tipo spray de elétrons e um analisador quadrupolo acoplado a um analisador TOF. O TOF é utilizado por apresentar vantagens como alta resolução e proporcionar leituras rápidas com massas acuradas.

Exemplo 4: Ionização por spray de elétrons:

Durante a ionização por spray de elétrons o analito é adicionado a uma solução com eletrólito em baixa concentração (~0.5%). A adição do analito a esta solução pode ser realizada de duas maneiras distintas: uma primeira maneira por meio de um sistema cromatográfico de micro ou nanofluxo e uma segunda maneira, por meio da injeção direta com o auxílio de uma seringa acoplada a uma bomba de micro

ou nanofluxo. Na presente concretização o meio de adição utilizado do analito a solução com o eletrólito foi à injeção direta do analito por meio de uma seringa acoplada a bomba de micro ou nanofluxo.

A solução analito-solvente é constantemente bombeada em uma extremidade do capilar, de modo a formar um menisco, o qual torna-se um "spray eletrolítico". A solução analito-solvente percorre por via de um capilar para o interior da câmera de ionização. Uma voltagem a qual varia na faixa de 2,3 a 5 kV é aplicada entre a extremidade do capilar contida no interior da câmera de ionização e o orifício de entrada do espectrômetro. Esta diferença de potencial faz a extremidade do capilar tornar-se um dos dipolos do campo elétrico gerado.

Este campo, pode alcançar até 10^6 V/m desestabiliza o menisco formado na extremidade do capilar por concentrar íons positivos (caso o espectrômetro de massa esteja em operação no modo positivo).

Um gás cônico é aplicado ao sistema a fim de acelerar a transição do estado físico do analito para que este possa ser analisado no espectrômetro de massa. O gás preferencialmente utilizado foi o gás N_2, a uma vazão de aproximadamente 80 l/h e a uma temperatura aproximada de 80°C na fonte. O gás acelera a evaporação do solvente das microgotas provenientes do spray, no qual as proteínas estão contidas.

Durante o processo de transição do estado líquido para o estado gasoso, as microgotas sofrem uma fissão de modo a tornarem-se gotículas altamente ionizadas. A repetição do

processo evaporação / fissão ocorrerá inúmeras vezes em processo simultâneo durante o trajeto das microgotas no percurso compreendido entre a ponta do capilar e o orifício de entrada do espectrômetro separado aproximadamente por 2 cm.

Um gás nobre é utilizado na câmera de colisão do espectrômetro de massa de modo a favorecer a leitura dos analitos. Na presente concretização o gás nobre utilizado foi o argônio. A calibração externa do analisador é preferencialmente realizada com iodeto de sódio sob uma extensão de massa preferencial no intervalo de 400 a 3000 m/z. Todos os espectros de massa foram obtidos com o analisador TOF em "V-mode" (TOF kV = 9,1) e voltagem MCP no quadro de 2,15kV.

Exemplo 5: Resultado da leitura em espectrometro de massa

Cada uma das amostras de soro foi injetada pelo menos duas vezes no espectrômetro de massa por meio de uma seringa, a qual estava acoplada ao dispositivo receptor da fonte com uma velocidade de escoamento de 1µL/min durante cerca de 2 min utilizando-se o módulo MCA do analisador TOF. No intervalo entre uma primeira injeção de amostras de soro e uma segunda amostra de soro, todo o sistema deve ser lavado com uma solução adequada, como por exemplo, uma solução de acetonitrila. Os dados para análise foram coletados no intervalo preferencial do espectro compreendido entre 400 a 3000 m/z.

Para os dados de espectometria de massa no intervalo aproximado de 1200 a 2200 m/z, os dados sofreram um tratamento computacional no programa Masslynx 3. Dito

programa computacional, aplica um "smooth filter" para reduzir ruídos. O "smooth filter" foi aplicado em 3 janelas do canal de modo a se utilizar o método da presente invenção.

O espectro multi-carga foi então convertido em um espectro de carga simples para o intervalo de 8 kDa a 250 kDa com o uso de um algoritmo de entropia máxima, o qual pertence ao programa computacional Masslynx. Entretanto, outros programas de desconvolução utilizando abordagem computacional parecida podem ser utlizados, não sendo limitado o uso do programa utilizado na presente invenção.

Para esse espectro, foi configurado uma resolução preferencial de 35 Da/canal com um modelo de dano de cerca de 0,75Da com a metade da largura da altura, raios de intensidade mínima para a esquerda e direita de aproximadamente 65%. Os dados de Massa/Intensidade foram exportados para os arquivos de texto.

Para uma segunda preparação no intervalo aproximado de 400 a 1200 m/z, os dados sofreram o mesmo tratamento computacional utilizado na leitura do espectro no intervalo aproximado de 1200 a 2200 m/z, ou seja, o programa computacional, o qual aplica um "smooth filter" para reduzir ruídos.

A prática da subtração com o uso de um polinômio na ordem de 40 graus e cerca da remoção de 30% sobre a curva de tolerância aproximada de 0, 010 foi aplicada para minimizar os efeitos sonoros.

Os dados de Massa/Intensidade foram exportados para arquivos de texto no formato ASCII (.txt) com a resolução

dos picos, de maneira a atingir a terceira casa decimal do Dalton de precisão.

Exemplo 6: Tratamento dos dados obtidos na leitura do espectro:

Os dados obtidos após o tratamento das leituras do espectro, foram analisados por meio do uso da estratégia SVM, a qual pode ser descrita conforme abaixo (Vapnik, V. N, 1995):

Dada um conjunto de treinamento linearmente separável no espaço de características: $S = \{(x_1, y_1), \ldots (x_n, y_n)\}$ que resulta na equação de um classificador linear $W^T x + b = 0$, onde **w** é o vetor normal e **b** um valor atribuído a um viéis, para uma amostra desconhecida com vetor de entrada **x**, este deve ser classificado com +1 se: $\langle w,x \rangle + b >= 1$ e classificado como -1 se: $\langle w,x \rangle + b <= -1$.

A Figura 1, mostra geometricamente, que a margem pode ser calculada de acordo com o desenvolvimento das etapas abaixo após a determinação do vetor normal:

$\langle w.x1 \rangle + b = 1$ (1.1)

$\langle w.x2 \rangle + b = -1$ (1.2)

Subtraindo 1.1 de 1.2

$w\langle x1 - x2 \rangle = 2$ (1.3)

Projetando o vetor de diferença sobre o vetor normal **w**:

$$\frac{1}{\|w\|} w \bullet \langle x1 - x2 \rangle = \frac{2}{\|w\|} \qquad (1.4)$$

O algoritmo busca o espaço de **w's** e **b's** de modo a encontrar a máxima margem de separação para posicionar um hiperplano. A melhor abordagem na resolução deste problema

é transforma-lo em um problema convexo, de modo a minimizar uma função quadrática sob restrições de inequações. Sendo assim, dito problema poderá ser solucionado, em sua forma dual, por meio da aplicação do tratamento de Lagrange.

$$L(w,b,\alpha)=\frac{1}{2}<w^T \bullet w>-\sum_{i=1}^{i=n}\alpha_i[y_i(<w \bullet x_i>+b)-1] \quad (1.5)$$

onde: $\alpha_i \ni 0$ são as variáveis na sua forma dual, ou multiplicadores de Lagrange.

A solução deste problema é equivalente à resolução da equação acima em sua forma dual (Wolfe), escrita somente como função das variáveis duais.

$$\text{Min} \quad L(\alpha) = \sum_{i=1}^{i=n}\alpha_i - \frac{1}{2}\sum_{j=1}^{j=l} y_i y_j \alpha_i \alpha_j <x_i \bullet x_j> \qquad (1.6)$$

Sujeito a: $\sum_i \alpha_i y_i = 0$ e $\alpha_i \geq 0$.

O vetor normal é obtido por meio da solucão este problema para os valores de α^*

$$w^* = \sum_i \alpha_i^* y_i x_i, \text{ for } \alpha_i^* > 0. \quad (1.7)$$

Para obter-se a função discriminante, f(x) = w.x + b o parâmetro de inclinação, ***b***, precisa ser computado. Isto é facilmente encontrado aplicando a condição "complementar" de Karush-Kuhn-Tucker:

$$\alpha_i \ (y_i(<w,x_i>+b) - 1) = 0. \quad (1.8)$$

A condição acima, somente é verificada para valores positivos de α_i. Esses multiplicadores estão associados aos pontos que definem a posição do hiperplano, sendo assim chamados de de vetores de suporte. Desta maneira, se o parâmetro de inclinação é corretamente computado, têm-se que para $\alpha_i > 0$, y_i ($\langle w, x_i \rangle + b$) = 1, de modo a satisfazer a equação de "complementaridade".

Uma abordagem para dados não separáveis pode ser por meio da utilização de *"slack variables"* (ξ) e / ou aplicação de funções kernel de forma não linear ($\emptyset$). Sendo assim o problema de otimização torna-se:

$$y_i((w\phi(x_i))+b \geq 1-\xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n \qquad (1.9)$$

O modelo permite alguns erros durante o processo de classificação, de forma que uma nova função passa a ser otimizada e:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \qquad (1.10)$$

onde C e uma constante > 0 e esta relacionada com o compromisso entre o erro empírico e a complexidade do modelo. A nova formulação torna-se

$$L_D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \phi(x_i x_j) \qquad (1.11)$$

sujeito a:

$$0 \leq \alpha_i \leq C, i = 1,\ldots n \qquad (1.12)$$

e também:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (1.13)$$

Ao se introduzir as *"slack-variables"*, limita-se o valor dos multiplicadores de Lagrange a um máximo de C ($\alpha_i \leq C$).

Exemplo 7: Preparação dos dados de SVM

O software "ACESO" (navegador sob o conjunto de espectro), desenvolvido no presente trabalho, foi utilizado para normalizar a intensidade dos espectros para valores entre 0 a 1 de tendo com resultado da corrente ionica máxima, o valor de 1. apropriados para aplicação do algoritmo. Adicionalmente, um valor médio para os dados do espectro é criado baseados nos dados do espectro de massa, multiplicados para cada amostra.

Para os espectros de peptideos (aproximadamente 400 - 1,200 m/z), o software configurou os dados do espectro de modo a possuírem cerca de 1 Da de resolução integrando os valores intermediarios. Desta forma, o softawer "ACESO" propriamente formado por dados em uma maneira otimizada para classificar e interagir com a próxima etapa com SVMPP, para classificar as informações baseadas nas aproximações do "leave one out".

As análises de "leave one out", foram realizadas para todas as amostras de soro dos 30 pacientes com a Doença de Hodgkin e para os 30 pacientes controles. As análises de "leave one out" são realizadas por meio da exclusão de um dado do arquivo para o assunto investigado e os restantes dos dados são utilizados como um quadro de treinamento.

O algoritmo da presente invenção usa pequenas porções dos espectros como conjunto de treinamento, de modo a buscar as regiões onde se obtém uma maior exatidão.

A exatidão do método é calculada como as funções positivas verdadeiras (TP), negativas verdadeiras (TN) falsos positivos (FP) e falsos negativos (FN) como mostrado na equação:

Exatidão = (TP + TN) / (TP + TN + FP + FN) .

Exemplo 8: Obtenção dos biomarcadores

O software "ACESO" em um segundo momento foi utilizado para promover a busca por biomarcadores por meio da análise de um pequeno pré-quadro "janela de estudo". A janela de estudo é uma pequena extensão em m/z cuja abertura é definida pelo usuário.

Duas análises de "leave one out" distintas foram realizadas para todas as janelas de estudos de modo a permanecer contra 59 quadros de assunto treinados na mesma extensão da janela; um primeiro grupo para as amostras de soro controle e um segundo grupo para as amostras de soro de pacientes contendo a Doença de Hodgkin.

A análise de MDA utilizou a janela para os valores aproximados de 100 m/z, 20 m/z do espectro e valores aproximados de 10 m/z do espectro para aproximadamente 400 a 1200 m/z de extensão e de aproximadamente 2,240 e 4,480 para 8 kDa a cerca de 200 kDa de extensão.

A produção dos dados de MDA é dada pelo arquivo texto do relatório para classificar todos as entradas de todas as janelas de estudo, e um gráfico no qual o distanciamento ordinário para os valores aproximados de 0 a 100 representa

a porcentagem de "material saudável" classificado em cada análise de "leave one out".

A abscissa do gráfico teve sua extensão analisada conforme a extensão dos dados obtidos no espectro total. Os dados de cada uma das análises de "leave one out" relativos a cada um dos grupos analisados foram plotados e conectados de modo a formar um atalho, o qual é mostrado na abscissa do gráfico.

O gráfico dos dados de MDA contém duas linhas paralelas no eixo x, aonde em um caso ideal, o primeiro cruzamento no eixo y em 100% e o segundo cruzamento no eixo y em 0%. A linha superior deve representar o grupo das amostras de sangue dos pacientes controles, de modo a indicar que cerca de 100% dos pacientes controles foram classificados com "saudáveis".

A linha inferior deve representar o grupo das amostras de sangue dos pacientes com a Doença de Hodgkin, ou seja, os pacientes "não-saudaveis", de modo a indicar que 0% dos pacientes deste grupo foram classificados como "saudáveis".

Entretanto, em uma base de dados atual não é esperado que ocorra este resultado. Pontos de convergência máxima entre as duas retas do gráfico, devem aparecer, de forma a representar a porção do espectro aonde a maior parte das amostras de sangue de pacientes controles e de amostras de sangue de pacientes infectados com a Doença de Hodgkin foram "corretamente" classificados.

Esses "hot spots", indicam regiões no gráfico aonde a busca por picos diferencialmente expressos no espectro para a representação de biomarcadores é ideal. Por esta razão a

técnica de extensão SVM foi nomeada como Análise de Divergêngia Máxima (MDA).

O algoritmo utilizado para o mecanismo de vetor de suporte foi capaz de classificar aproximadamente 93% das amostras de sangue dos pacientes controles e aproximadamente 88% das amostras dos pacientes com a Doença de Hodgkin pela técnica "leave one out" com uma exatidão de aproximadamente 90%.

As amostras de sangue dos pacientes controles foram classificadas como pertencentes à classe saudável ou a classe doente. Os pacientes acometidos pela Doença de Hodgkin que foram classificados erroneamente são os pacientes: 4, 5, 16, 20. As amostras de soro identificadas como 5, 16 e 20 são de pacientes HIV+.

As chances de se ter selecionado 4 pacientes, dentre os quais 3 ou mais são pacientes HIV+ para uma população de 30 amostras de sangue de pacientes, a qual já continha cerca de 6 pacientes HIV+ é menor que 1%. Este fato indica que a infecção causada pelo HIV proporciona uma alteração na proteína associada ao espectro de massa para pacientes infectados com a doença de Hodgkin.

Dentre o grupo de pacientes com Doença de Hodgkin, o paciente 4 além de mostrar um teste negativo de imuno-histoquimica para o vírus EBV, mostrou que o estágio de progressão da Doença de Hodgkin era inicial, fato este, que pode ser explicado pela classificação incorreta.

Esta metodologia pode ser estendida para a criação de outros modelos como, por exemplo, o diagnóstico múltiplo. O presente método de sistema de diagnóstico baseado na

técnica SVM poder ser utilizado para diagnosticar populações, as quais contém pacientes com DH e pacientes com DH + HIV.

Em um segundo intervalo de leitura do espectro 400 a 1200m/z, o algoritmo do mecanismo de vetor de suporte classificou todas as matérias controles e os pacientes com a Doença de Hodgkin "corretamente" por meio da técnica de "leave one out". Este resultado mostra que para esse intervalo do espectro, os dados obtidos indicam que a extensão de aproximadamente 400 a 1200 m/z é a extensão mais recomendada na utilização de classificação para a Doença de Hodgkin associada a outras patologias do que os dados de alta massa molecular, relacionadas no estado da arte.

A Figure 2 mostra os resultado das análises MDA com o uso de uma abertura da janela de estudos de aproximadamente 2240 e 4480 Da. As análises para a janela de estudo de aproximadamente 4480 Da mostra uma importante região de divergência em torno da zona de aproximadamente 130 kDa.

Em aproximadamente 2240 Da o resultado da análise de MDA confirma este segmento chave aproximadamente entre os valores de 131 kDa e 133 kDa de modo a apresentar uma ótima divergência. As análises de MDA nesta região para todas as amostras de soro dos pacientes controles e dos pacientes com a Doença de Hodgkin expressam picos diferenciais de aproximadamente 132, 740Da, 97% de pacientes com a Doença de Hodgkin e 97% para o soro das amostras de sangue dos pacientes-controle, esses picos não são expressos.

Os pacientes de numero 305 e novamente 16 forma classificados erroneamente no outro ponto de máxima divergência entre os espectros.

Um eficiente caminho para examinar muitos espectros em busca de biomarcadores é a comparação da média de todos os espectros correspondentes a cada pico, para cada classe.

A média do espectro foi construída pela determinação da média da intensidade de massa de cada pico para cada um dos grupos. O espectro de massa para esta região é mostrado na Figura 3, a presença de pico expresso em aproximadamente 132, 740Da para amostras de sangue de pacientes controles é diferentemente expresso para amostras de sangue de pacientes com a Doença de Hodgkin.

As análises de MDA para a janela de estudo de aproximadamente 20 m/z e 10 m/z mostram claramente a divergência do segmento de extensão de aproximadamente 980 m/z a 1000 m/z com máxima divergência aproximada entre 990 m/z - 1000 m/z, conforme mostra a Figura 4. Esta região do espectro indica o sítio indicativo de potenciais biomarcadores para diagnóstico clínico.

De acordo com a Figura 5, O espectro de massa para a região de aproximadamente 980 m/z mostra a presença de envelopes isotópicos expressos diferentemente em aproximadamente 980 e 994 m/z em amostras de sangue de pacientes controles. Ditos envelopes isotópicos não são expressos no espectro de amostras de pacientes com a Doença de Hodgikin.

Pela performance da aproximação das análises de "leave one out" somente sob o segmento entre cerca de 990 e 1,000

Da, aproximadamente 97% dos pacientes acometidos pela Doença de Hodgkin foram "corretamente" classificados assim como aproximadamente 91% das amostras de sangue de pacientes controles, conforme mostra a Figura 5. As amostras de sangue de pacientes controles pela classificação errônea são 5 e 299. Uma classificação errônea das amostras de sangue dos pacientes # 9 com a Doença de Hodgkin mostrou uma imunohistoquimica negativa para o vírus EBV.

Para promover o estudo das amostras do material controle 5 e 299, foi realizado o teste de PCR para o vírus EBV, no qual foram confirmados os resultados positivos mostrados para ambas as amostras de soro dos pacientes controles 5 e 299. Um grande número de pacientes com a Doença de Hodgkin também apresentou uma alta taxa do anticorpo anti EBV em seus soros.

A classificação "não esperada" dos pacientes 5 e 299, ocorreu devido a ao fato desses pacientes apresentarem uma alta taxa de EBV presente em seus soros. O modelo proposto na presente invenção foi treinado tomando como base pacientes com HD, os quais também possuíam o vírus EBV. Desta forma, a presença do vírus EBV em seus soros foi detectada, de modo a classificar esses pacientes de uma maneira errônea.

A evolução das extensas massas espectrométricas tanto para as amostras de sangue dos pacientes com a Doença de Hodgkin quanto para as amostras de sangue dos pacientes "saudáveis", tido como matéria controle, mostram que o modelo foi capaz de classificar individuos com maior

exatidão (400 - 1,200 m/z), do que na faixa de maior massa molecular.

Nos dados do espectro estendido para aproximadamente 1200-2200 m/z, conforme mostra a Figura 4, um bom índice de acertos no processo de classificação foi alcançado, mas o método foi capaz de apontar "hot spots" no espectro. Estes "hot spots" são capazes de separar os resultados obtidos entre o grupo de pacientes controles e o grupo de pacientes com HD além de discriminar padrões oriundos do vírus HIV e do vírus EBV.

A análise de MDA pode ser interpretada como um aspecto de seleção, e cada aspecto isolado representa um novo biomarcador para diagnóstico médico. Na presente invenção, cerca de 100% das amostras de matéria controle e das amostras de pacientes infectados com a Doença de Hodgkin foram corretamente classificados na extensão aproximada de 400 - 1200 m/z.

Através do método desenvolvido pela presente invenção é possível o diagnóstico rápido do câncer, de modo a permitir assim um tratamento personalizado.

A invenção descrita, assim como os aspectos abordados devem ser considerados como possíveis concretizações. Deve, entretanto ficar claro que a invenção não está limitada a essas concretizações e, aqueles com habilidade na arte irão perceber que, qualquer característica particular nela introduzida, deve ser entendida apenas como algo que foi descrito para facilitar a compreensão. As características limitantes do objeto da invenção estão relacionadas às reivindicações que fazem parte do presente relatório.

REIVINDICAÇÕES

1. Método de diagnóstico baseado em padrões proteômicos e/ou genômicos por meio da análise de SVM **caracterizado por** buscar preferencialmente um pequeno perfil proteômico expresso por meio dos picos do espectro de espectrometria, utilizando a técnica de espectrometria de massa em diferentes intervalos do espectro.

2. Método de diagnóstico de acordo com a reivindicação 1, **caracterizado por** utilizar a metodologia de máquinas de vetores de suporte para classificar uma amostra como de doente ou saudável, baseado em todo ou em parte do perfil proteômico obtido no espectrômetro de massas.

3. Método de diagnóstico de acordo com a reivindicação 1, **caracterizado por** os dados da análise compreendida entre o intervalo aproximado de 1200 a 2200 m/z e 400 a 1200 m/z, sofrerem um tratamento computacional no programa Masslynx 3 ou similar.

4. Método de diagnóstico de acordo com a reivindicação 1, **caracterizado por** os dados das leituras do espectro serem analisados por meio do uso da estratégia SVM, serve para obter a máxima margem de separação para posicionar um hiperplano.

5. Método de diagnóstico de acordo com a reivindicação 4, **caracterizado por** a abordagem para dados não separáveis ser por meio da utilização de "*slack variables*" (ξ) e / ou aplicação de funções kernel de forma não linear ($\varnothing$).

6. Método de diagnóstico **caracterizado por** os dados obtidos na análise de SVM serem tratados por meio de um programa de computador, dito programa utilizado para: (i)

normalizar a intensidade dos espectros para valores entre 0 a 1, tendo com resultado da corrente iônica máxima, o valor de 1; e, (ii) classificar e interagir com a etapa SVMPP, para classificar as informações baseadas nas aproximações do "leave one out".

7. Método de diagnóstico de acordo com a reivindicação 6, **caracterizado por** para os espectros de peptideos (aproximadamente 400 - 1,200 m/z), o programa de computador configurar os dados do espectro de modo a possuírem cerca de 1 Da de resolução integrando os valores intermediários.

8. Método de obtenção de biomarcadores para diagnóstico por meio de programa de computador, **caracterizado por** utilizar análises de um pequeno pré-quadro "janela de estudo" de pequena extensão em m/z cuja abertura é definida pelo usuário.

9. Método de acordo com a reivindicação 8, **caracterizado por** a produção dos dados ser dado pelo arquivo texto do relatório para classificar todos as entradas de todas as janelas de estudo, e um gráfico no qual o distanciamento ordinário para os valores aproximados de 0 a 100 representa a porcentagem de "material saudável" classificado em cada análise de "leave one out".

10. Método de acordo com a reivindicação 9, **caracterizado por** o gráfico conter uma linha superior no eixo x, representa o grupo das amostras de sangue dos pacientes controles classificados como "saudáveis", uma linha inferior no eixo x, representa o grupo das amostras de sangue dos pacientes com a Doença de Hodgkin, os pacientes "não-saudaveis".

11. Método de acordo com a reivindicação 9, **caracterizado por** o gráfico apresentar pontos de convergência máxima entre as duas retas, representa a porção do espectro aonde a maior parte das amostras de sangue foi "corretamente" classificada, indicando o sítio de potenciais biomarcadores para diagnóstico clínico.

12. Método de acordo com a reivindicação 9, **caracterizado por** a metodologia do programa de computador ser utilizada ainda para o diagnostico de outras doenças.

13. Biomarcadores caracterizado por serem determinados através das análises de SVM, após localização das janelas de interesse e posteriormente após localização através do espectro de massa, de modo que a identificação dos ditos biomarcadores ocorra por meio de um gel 2D ou por meio da espectrometria de massa.

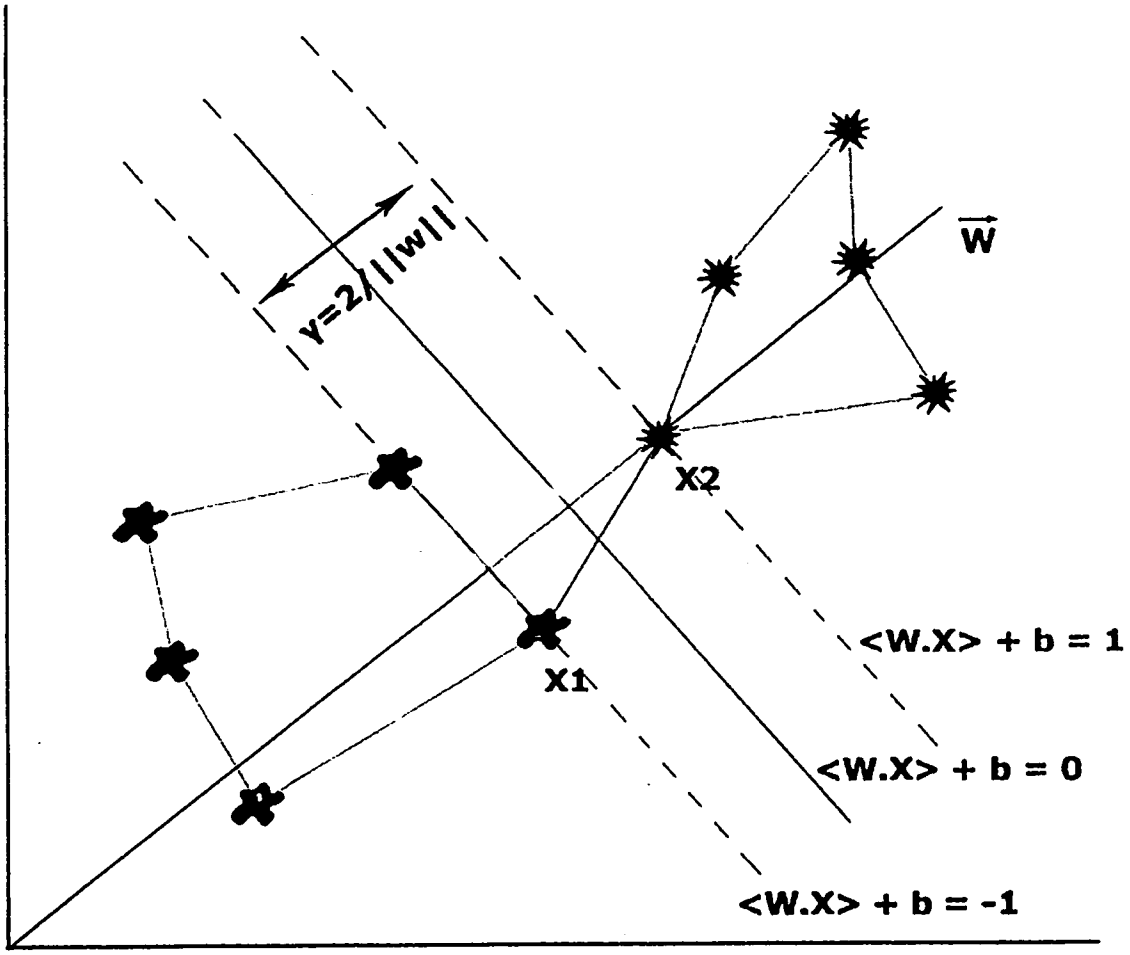


FIGURA 1

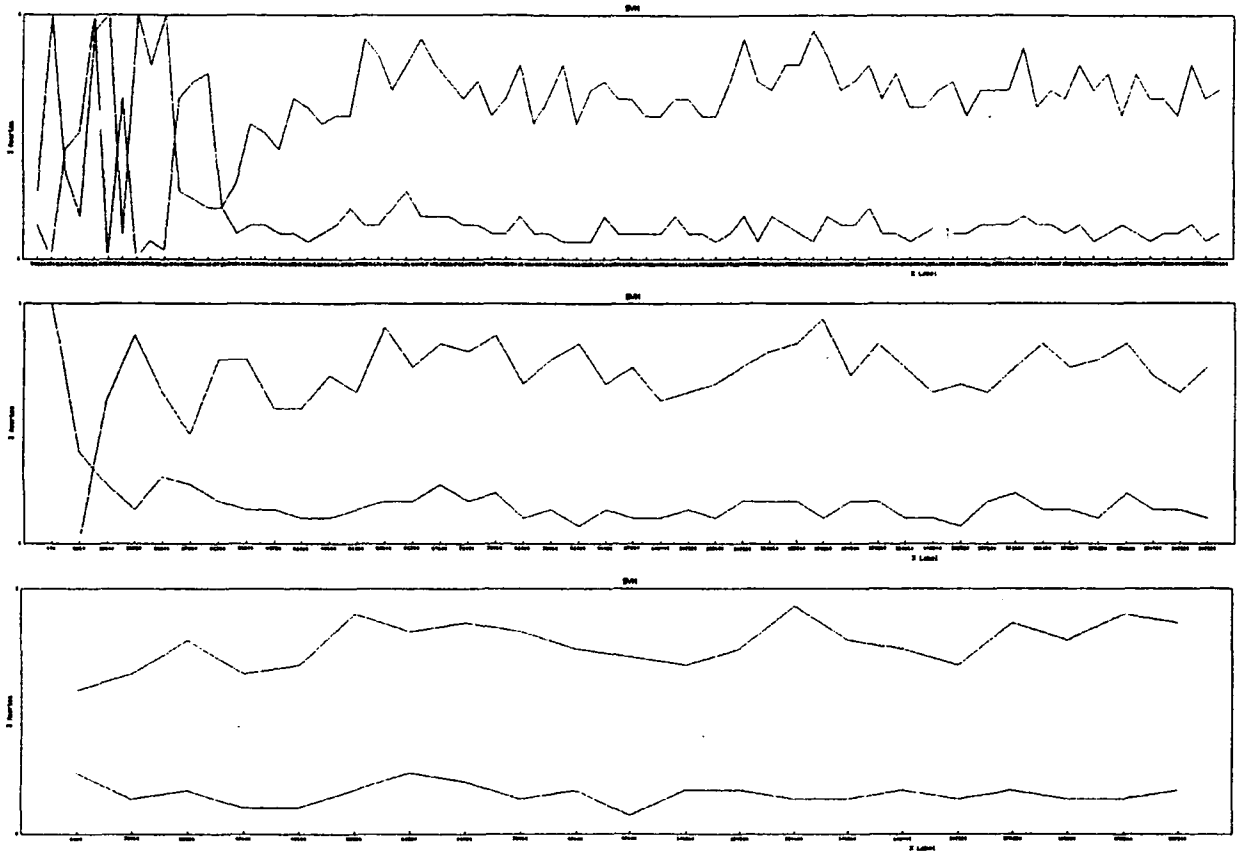


FIGURA 2

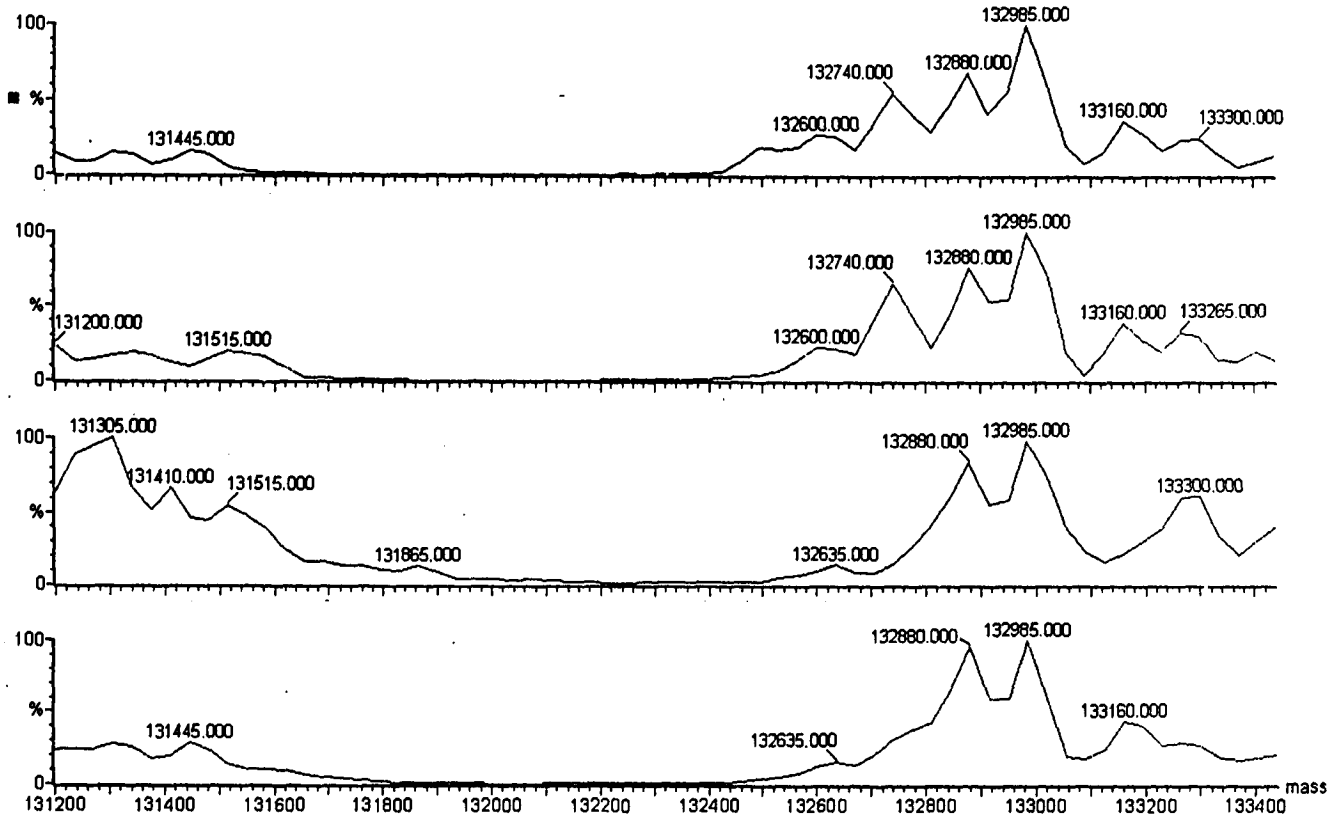

100
%
0
131445.000
132600.000
132740.000
132880.000
132985.000
133160.000
133300.000
100
%
0
131200.000
131515.000
132600.000
132740.000
132880.000
132985.000
133160.000
133265.000
100
%
0
131305.000
131410.000
131515.000
131865.000
132635.000
132880.000
132985.000
133300.000
100
%
0
131445.000
132635.000
132880.000
132985.000
133160.000
131200
131400
131600
131800
132000
132200
132400
132600
132800
133000
133200
133400
mass

FIGURA 3

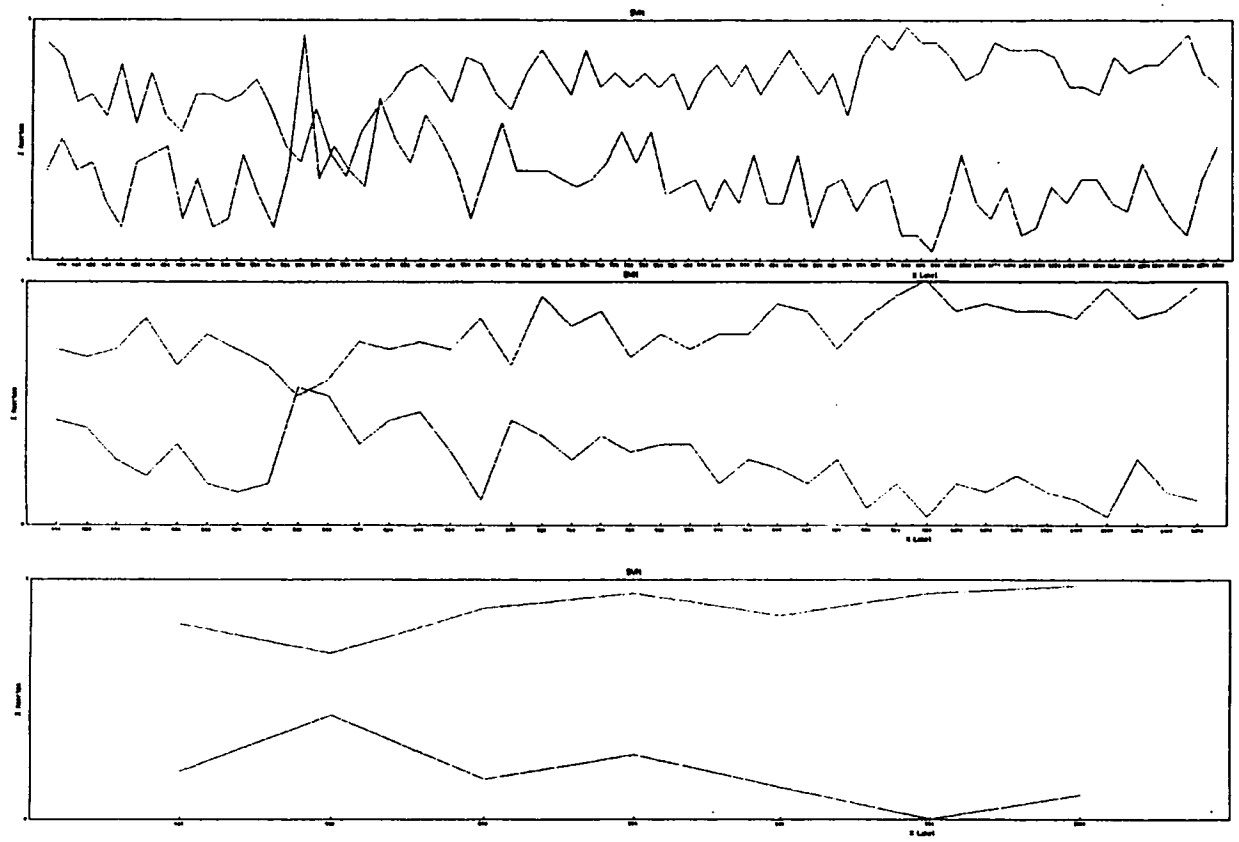


FIGURA 4

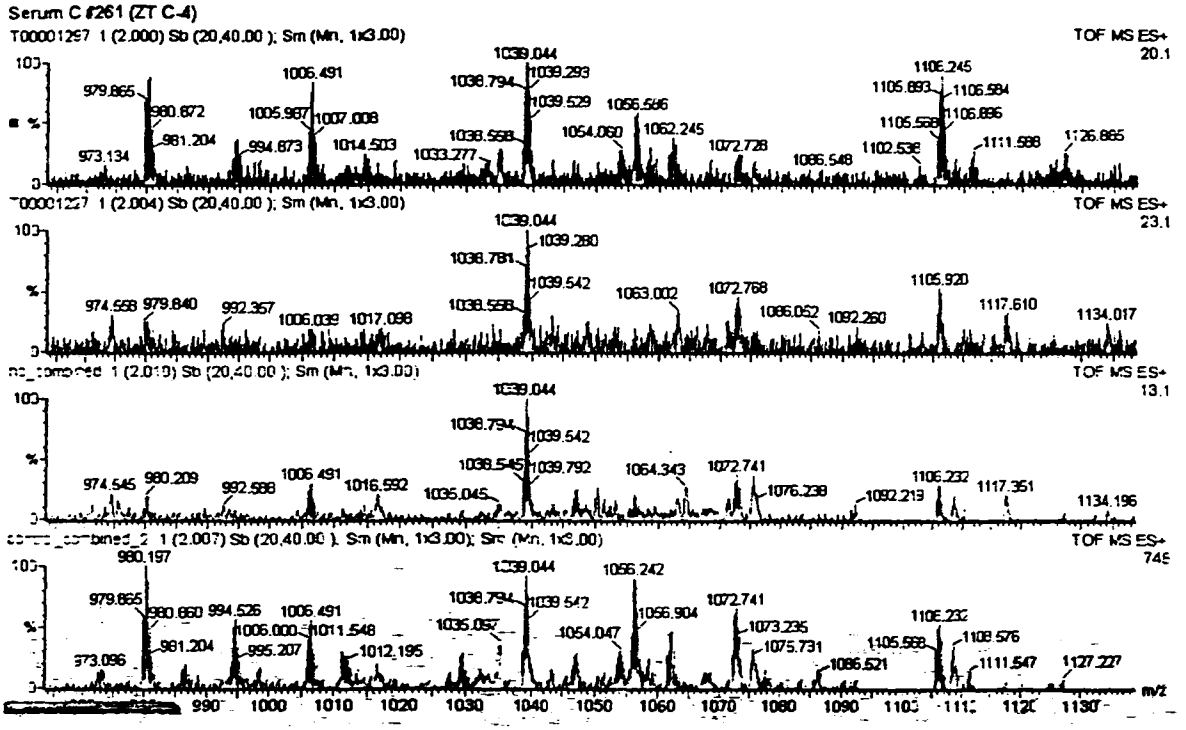

Serum C #261 (ZT C-4)
TOF MS ES+
20.1
1039.044
1039.293
1038.794
1039.529
1056.566
1006.491
979.865
980.672
1005.987
1007.008
981.204
994.873
1014.503
973.134
1033.277
1038.558
1054.060
1062.245
1072.728
1086.548
1102.538
1105.245
1105.893
1106.584
1105.558
1106.896
1111.598
1126.865
TOF MS ES+
23.1
1039.044
1039.280
1038.781
1039.542
1038.558
1063.002
1072.768
1086.052
1092.260
1105.920
1117.610
1134.017
974.568
979.840
992.357
1006.039
1017.098
TOF MS ES+
13.1
1039.044
1038.794
1039.542
1038.545
1039.792
1064.343
1072.741
1076.238
1092.219
1106.232
1117.351
1134.196
974.545
980.209
992.588
1006.491
1016.592
1035.045
TOF MS ES+
745
980.197
1039.044
1056.242
979.865
980.660
994.526
1006.491
1038.794
1039.542
1072.741
1006.000
1011.548
1056.904
1035.057
1073.235
1106.232
981.204
995.207
1054.047
1075.731
1105.568
1108.576
973.096
1012.195
1086.521
1111.547
1127.227
990
1000
1010
1020
1030
1040
1050
1060
1070
1080
1090
1100
1110
1120
1130
m/z

Figura 5

RESUMO

Método de diagnóstico baseado em padrões proteômicos e/ou genômicos por vetores de suporte aplicado a espectometria de massa

A presente invenção refere-se a um método de diagnóstico médico baseado em padrões proteômicos e/ou genômicos, por meio de dados obtidos em espectros gerados pela espectometria de massa no qual é possível classificar os pacientes quanto ao estágio de infectabilidade. Adicionalmente, a presente invenção também se refere a dois novos biomarcadores para o diagnóstico médico da Doença de Hodgkin. Baseado nas análises de SVM localiza-se as janelas de interesse e posteriormente recorre-se ao espectro de massa para que ocorra a localização dos biomarcadores, de modo que a identificação dos ditos biomarcadores ocorra por meio de um gel 2D ou por meio da espectrometria de massa.