Universidade Federal do Maranhão

Centro de Ciências Exatas e Tecnologia

Programa de Pós Graduação em Engenharia de Eletricidade

Simone Cristina Ferreira Neves

Classificação de câncer de ovário através de padrão proteômico e análise de componentes independentes

São Luís

2012

Universidade Federal do Maranhão

Centro de Ciências Exatas e Tecnologia

Programa de Pós Graduação em Engenharia de Eletricidade

Simone Cristina Ferreira Neves

Classificação de câncer de ovário através de padrão proteômico e análise de componentes independentes

Dissertação apresentada ao curso de
Pós-Graduação em Engenharia de Eletricidade da
UFMA como parte dos requisitos necessários para
obtenção do grau de mestre em Engenharia Elétrica.

São Luís

2012

Neves, Simone Cristina Ferreira.

Classificação de câncer de ovário atravésde padrão proteômico e análise de componentes independentes/Simone Cristina Ferreira Neves – São Luís, 2012.

56 f.

Impresso por computador (fotocópia).

Orientador: Allan Kardec Duailibe Barros Filho.

Dissertação (Mestrado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2012.

1. Padrão proteômico - Câncer de ovário. 2. Análise de componentes independentes. 3. Redes neurais. I. Título.

CDU 004:618.11-006

CLASSIFICAÇÃO DE CÂNCER DE OVÁRIO ATRAVÉS DE PADRÕES PROTEÔMICOS E ANÁLISE DE COMPONENTES INDEPENDENTES

Simone Cristina Ferreira Neves

Dissertação aprovada em 24 de julho de 2012.

Prof. Allan Kardec Duailibe Barros Filho, PhD.
(Orientador)

Prof. Ewaldo Eder Carvalho Santana, Dr.
(Co-orientador)

Prof. Vicente Silva Gonçalves Neto, Dr.
(Membro da Banca Examinadora)

Prof. Sofiane Labidi, Dr.
(Membro da Banca Examinadora)

À Deus, pela vida.

Ao Rômulo, meu esposo e aos meus pais pela compreensão e carinho.

Aos amigos pela existência, apoio e incentivo.

Ao professor Allan Kardec pela oportunidade e aprendizado.

E a UFMA.

Agradecimentos

Ao Senhor da minha vida, o Criador de todas as coisas;

Aos professores Dr. Allan Kardec B. Dualibe Filho pela oportunidade, apoio, dedicação, competência e paciência e ao professor Dr. Ewaldo Santana e Dr. Maria da Guia pelo apoio e dedicação.

Ao meu esposo Rômulo, aos meus pais Neves e Helena, aos meus irmãos Cláudia, Thallysson, cunhados e toda família, que sempre estiveram ao meu lado, agradeço pelo amor e paciência;

Aos meus amigos do laboratório de Processamento da informação biológica (PIB): Áurea Celeste, Cristiane, Denner, Eder Junior, Flávio Mello, Lúcio Flávio, Sidcley, Enio Aguiar, Anderson Brito, Luís Henrique, Luís Oliveira, Marcos Vinícius, Thiago Alexandro, Dr. Orlando, Professor Vicente, Ana Lúcia e Geraldo Junior.

Aos meus grandes amigos, Pinheiro Moura e Marco Souza, por sempre estarem me ajudando em vários momentos.

A todos que de alguma forma contribuíram para este período de aprendizagem.

“A experiência é uma lanterna dependurada nas costas que apenas ilumina o caminho já percorrido.”

Confúcio

Resumo

O câncer de ovário possui difícil diagnóstico nas primeiras fases de desenvolvimento. Neste trabalho trazemos um estudo de um novo método que nos deu ótimas taxas de precisão baseado em uma ferramenta da bio-informática chamada superfície mehorada a laser para ionização e dessorção (SELDI-TOF) usada para geração de padrões proteômicos que é uma das tecnologias mais avançada no auxílio ao diagnóstico. Nosso objetivo é contribuir para eficácia desta esta ferramenta, que já auxilia o dignóstico precoce, nossa metodologia usa análise de componentes independentes (ICA) para extração de caractéristicas e redes neurais para classificar entre malignidade e não malignidade em uma base de dados do centro de pesquisa do câncer nos EUA. Nosso trabalho obteve taxas de 97% de acurária, 98% de especifidade e 96 % de sensibilidade.

Palavras-chave: Câncer de ovários, padrões proteômicos, análise de componentes principais, análise de componentes independentes e redes neurais artificiais.

ABstract

The ovarian cancer is difficult to diagnose in the early stages of development. In this work we bring a study of a new method that gave us great accuracy rates based on a bioinformatics tool called surface enhanced for laser desorption and ionization (SELDI-TOF) used to generate proteomic patterns which is one of the technologies advanced in the diagnosis. Our goal is to contribute to effectiveness of this tool, which already helps diagnosis earlier, our methodology uses independent component analysis (ICA) for feature extraction and neural networks to classify between malignancy and no malignancy in a database of the research center cancer in the U.S.A. Our work rates obtained acurracy 97%, 98% specificity and 96% sensitivity.

Keywords: ovarian cancer, proteomic patterns, principal component analysis, independent component analysis and artificial neural networks.

Lista de figuras

Lista de tabelas

Lista de siglas

INCA	Instituto nacional do câncer
ACS	American Cancer Society
ICA	Análise de componentes independentes
PCA	Análise de componentes principais
SELDI	Surface-enhanced laser desorption/ionization
MALDI	Matrix-assisted laser desorption/ionization
MS	Espectrometria de massa
TOF	Time of flight
RNA	Redes neurais artificiais
MLP	Multilayer Perceptron
FBR	Função de base radial

SUMÁRIO

Capítulo 1

Introdução

O câncer de ovário é um tumor ginecológico que possui diagnóstico difícil na fase inicial e geralmente quando é diagnosticado já está em fase mais avançada, neste momento a chance de cura diminui drasticamente. Um dos problemas é que mesmo em países desenvolvidos existem poucos avanços no diagnóstico precoce, o que continua colaborando para que a taxa de mortalidade seja elevada (1)

A estimativa de novos casos em 2012 é de 6.190 e o número de morte em 2009 com câncer de ovário foi de 2.963 no Brasil(1), e de acordo com (2) nos Estados Unidos a estimativa para este ano é 22.280 novos casos, e 15.500 mortes.

Distribuição proporcional dos dez tipos de câncer mais incidentes estimados para 2012 por sexo, exceto pele não melanoma*

Localização primária	casos novos	percentual			Localização primária	casos novos	percentual
Próstata	60.180	30,8%	Homens	Mulheres	Mama Feminina	52.680	27,9%
Traqueia, Brônquio e Pulmão	17.210	8,8%			Colo do Útero	17.540	9,3%
Cólon e Reto	14.180	7,3%			Cólon e Reto	15.960	8,4%
Estômago	12.670	6,5%			Glândula Tireoide	10.590	5,6%
Cavidade Oral	9.990	5,1%			Traqueia, Brônquio e Pulmão	10.110	5,3%
Esôfago	7.770	4,0%			Estômago	7.420	3,9%
Bexiga	6.210	3,2%			Ovário	6.190	3,3%
Laringe	6.110	3,1%			Corpo do Útero	4.520	2,4%
Linfoma não Hodgkin	5.190	2,7%			Linfoma não Hodgkin	4.450	2,4%
Sistema Nervoso Central	4.820	2,5%			Sistema Nervoso Central	4.450	2,4%

*Números arredondados para 10 ou múltiplos de 10

Figura 1: Estimativas do câncer de ovário. Fonte: Adaptada de **(1)**.

O câncer de ovário é sétimo câncer mais comum entre as mulheres, de acordo com o instituto nacional do câncer - INCA. Ele ocupa o quinto lugar nas mortes por câncer entre as mulheres, sendo responsável por mais mortes do que

qualquer outro câncer do sistema reprodutivo feminino. O câncer de ovário é responsável por cerca de 3% de todos os cânceres nas mulheres. O risco de uma mulher desenvolver um câncer de ovário durante sua vida útil é de 1 em 71. Sua chance de morrer de câncer de ovário é de 1 em 95. Esse tipo de câncer se desenvolve principalmente em mulheres mais velhas, e cerca de metade das mulheres que são diagnosticadas com câncer de ovário têm 60 anos ou mais (3) (4).

Existem vários tipos de câncer de ovário, os três principais são o *epitelial (carcinomas),* este é o mais freqüente, o *tumor em células germinativas* e o *tumor em células estromais* (3).

O câncer de ovário epitelial pode acometer até mulheres jovens na faixa dos 15 anos, mas a idade média das pacientes é de aproximadamente 56 anos. Esta neoplasia acomete mais comumente as mulheres de etnia branca dos países industrializados da Europa Ocidental e da América do Norte, sendo menos comum na Índia e na Ásia, as mulheres destes locais têm um risco baixo, mas não quando vivem na América do Norte ou na Europa, as escandinavas e norueguesas têm o maior risco, em mulheres que não tiveram filhos o risco também é maior, nas que apresentaram menopausa precoce ou menopausa tardia, e nas mulheres com história de câncer de mama têm um maior risco de apresentar câncer (5).

O câncer de ovário pode ser também hereditário e este é o fator de risco mais relevante, história familiar da doença em um ou mais parentes de primeiro grau. O risco é um pouco menor para as mulheres com apenas um parente de primeiro grau ou um de segundo grau (avó ou tia) com câncer de ovário (5).

Na maioria das famílias afetadas com a síndrome de câncer de mama-ovário ou ovário isolado, um sítio específico de alteração genética foi encontrado para um o locus BRCA1 no cromossomo 17q21. O gene BRCA2, no cromossomo 13q12, também é responsável por alguns casos de câncer de ovário hereditário e câncer de mama (4).

O risco ao longo da vida para desenvolver câncer de ovário em pacientes portadores de mutações germinativas em BRCA1 é substancialmente maior sobre a

população em geral, motivo pelo qual essas pacientes merecem exames de rastreamento diferenciados (4).

1.1 O Diagnóstico do câncer de ovário

O câncer de ovário diferentemente de outros cânceres, como o câncer de mama, não permite a paciente um auto-exame, o que ajudaria muito na maioria dos casos pois geralmente o diagnóstico é tardio, nos estágios III e IV.

Para melhor esclarecimento o câncer de ovário apresenta os seguintes estágios e fases (6):

Estágio I – O crescimento é limitado aos ovários

Estágio II – O crescimento está envolvendo um ou dois ovários e com extensão para a região pelve.

Estágio III – O tumor já acomete um ou ambos os ovários, com implantes peritoneais comprovados além da região pelve; e/ou com metástases; tumor limitado à pelve, mas com propagação histologicamente comprovada iniciada em outros orgãos.

Estágio IV – O crescimento está envolvendo um ou ambos os ovários e com metástases à distância, ou seja, envolvendo outros órgãos ou todo o corpo(7).

A presença de um tumor é apontada por um exame normal de ultra-sonografia, mas a partir daí deve-se imediatamente buscar o diagnóstico diferencial de benignidade e malignidade dos tumores do ovário. Caso a suspeita de malignidade seja elevada, a paciente deverá ser encaminhada para centro de atendimento especializado para o tratamento adequado, por equipe capacitada para a abordagem dessa doença. No diagnóstico do especialista para formação de uma base de informações são utilizados métodos como a ultra-sonografia, marcadores específicos e informações da paciente como idade, tratamentos passados no sistema reprodutor e hereditariedade de câncer(8).

No Brasil o diagnóstico utiliza métodos que auxiliam o diagnóstico como ultrasonografia, e o CA 125, mas o diagnóstico só é certificado mediante a biopsia do tumor(9).

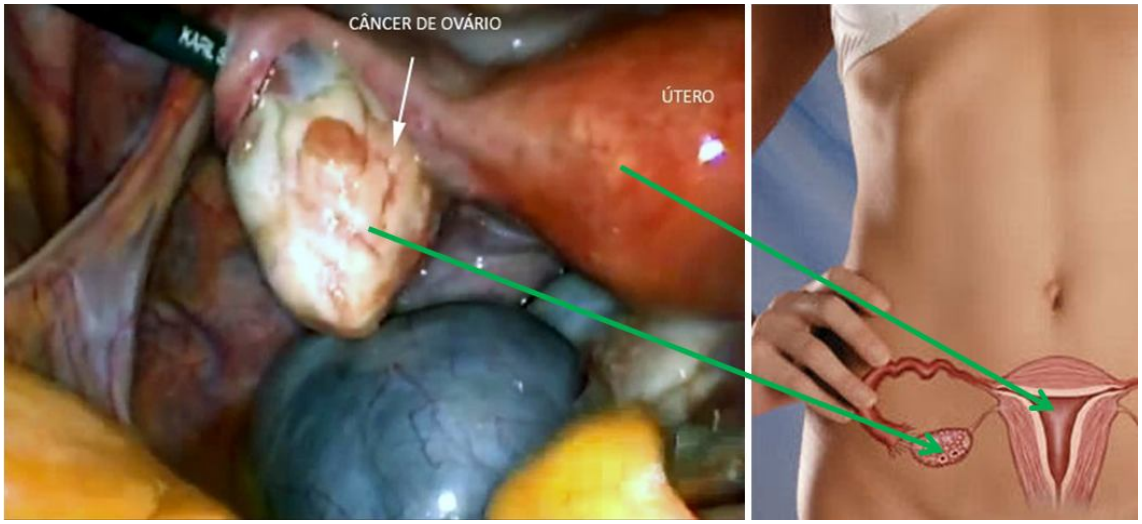

Figura 2: Câncer de ovário. Fonte: Adaptada de **(10)**.

A ultra-sonografia é o método propedêutico mais solicitado para o diagnóstico diferencial de tumores pélvicos. É um método com elevada precisão para a determinação de presença, tamanho, localização e caráter destes tumores. Vários estudos tem sido conduzidos no sentido de verificar alterações ultra-sonográficas sugestivas de malignidade. Quanto mais complexo à ultra-sonografia mostra-se o tumor ovariano, maior o risco de ser maligno. Vários aspectos ultra-sonográficos tem sido utilizados para caracterizar a benignidade ou a malignidade de tumores ovarianos: tamanho, multilocularidade, presença de partes sólidas, excrescências papilares, septos e suas características, alteração da ecogenicidade, bilateralidade, ascite e metástases (11).

O tamanho do tumor é importante, principalmente após a menopausa. Pesquisas em 36 mulheres na pós-menopausa portadoras de tumores pélvicos e observaram que a mediana do volume dos tumores benignos foi de 85,2 cm^3, e dos malignos, 452,5 cm^3 (11), este estudo mostra nove casos de malignidade (47,4%) em 19 portadoras de cistos ovarianos com septos finos e sete em oito (87,5%) portadoras de cistos ovarianos com septos espessos. A presença de áreas sólidas é também considerada critério de malignidade. É importante caracterizar o tipo de tumor sólido, pois muitas vezes o que é descrito como lesão complexa trata-se

apenas do resíduo ovariano ainda presente dentro da formação cística, os tumores benignos apresentavam-se com partes sólidas em 58% dos casos e os malignos em 93% (11).

Em seguida o CA-125, vem auxiliando no diagnóstico, este é um antígeno glicoprotéico de alto peso molecular descoberto por Robert Bast em 1981. Trata-se de um dos marcadores tumorais mais utilizados em oncologia ginecológica. Não é encontrado em ovário adulto normal, entretanto níveis elevados são encontrados no câncer de ovário, sendo usado como marcador tumoral, sua dosagem também pode ser detectada em outras condições, endometriose, gestação, pacientes saudáveis, câncer endométrico, e outras neoplasias. O CA-125 reage com as seis linhagens celulares das neoplasias malignas epiteliais do ovário, a saber: serosa, endometrióide, mucinosa, de células claras, indiferenciada e de histologia mista. Verificaram ainda que há maior freqüência de casos com valores superiores aos valores de corte quanto mais avançado o nível do tumor (11).

1.2 A bio-informática e sua contribuição para proteômica.

Podemos considerar a bioinformática como uma linha de pesquisa que envolve aspectos multidisciplinares e que surgiu a partir do momento em que se iniciou a utilização de ferramentas computacionais para a análise de dados genéticos, bioquímicos e de biologia molecular. A bioinformática envolve a união de diversas linhas de conhecimento – a ciência da computação, a engenharia de softwares, a matemática, a estatística e a biologia molecular – e tem como finalidade principal desvendar a grande quantidade de dados que vem sendo obtida através de seqüências de DNA e proteínas. Para o desenvolvimento de genomas completos, a informática é imprescindível e a biologia molecular moderna não estaria tão avançada hoje, não fossem os recursos computacionais existentes.

Desde que os seqüenciadores capilares de DNA em larga escala surgiram, no fim da década de 90, a quantidade de dados biológicos produzidas simplesmente alcançou níveis que fizeram com que análises manuais de seqüências de DNA se tornassem simplesmente alternativas absurdas para o estudo de dados do genoma. Dois desenvolvimentos foram importantes para permitir tanto o surgimento da bio-informática quanto o rápido desenvolvimento da produção de seqüências de DNA. O

primeiro deles foi o sequenciamento capilar e o outro grande desenvolvimento foi a marcação das moléculas necessárias para o sequenciamento do DNA. Enquanto as reações tradicionais eram realizadas com marcadores radioativos, que tornavam a metodologia um tanto quanto trabalhosa e até mesmo perigosa, os marcadores fluorescentes permitiam maior segurança e ainda um novo avanço. Enquanto era preciso ocorrer diferentes reações na marcação radioativa, a técnica de marcação fluorescente permitia que cada base fosse marcada separadamente e utilizava uma técnica capaz de emitir luz em um diferente comprimento de onda se excitado por um laser. Essa luz lida por um detector informava ao sistema dados precisos da eletroforese. E foi exatamente a reunião desses dois desenvolvimentos num só aparelho que produziu o equipamento que posteriormente ficaria conhecido como “o seqüenciador que criou a bioinformática”. O primeiro desses aparelhos foi produzido pela empresa Applied Biosystems e foi chamado de ABI Prism 3700 (12).

Após a conclusão do seqüenciamento dos genomas de vários organismos, inclusive o da espécie humana, a pesquisa envolvendo proteínas ganhou novo fôlego, e por isso a química das proteínas tem passado por grandes avanços tecnológicos, dando um grande impulso para o estudo das proteínas(8).

1.1.1 Proteoma

O nome proteoma se originou porque este indica as ***prote***ínas de um gen***oma*** ou tecido. Enquanto o genoma representa a soma de todos os genes de um indivíduo, que são características fixas, o proteoma não possui as características fixas de um organismo. O proteoma é alterado conforme o estado de desenvolvimento do tecido ou sob as condições nas quais o indivíduo se encontra. O proteoma nos ajuda a compreender molecularmente como uma célula funciona em um indivíduo doente e um indivíduo sadio, pois é preciso ter conhecimento das proteínas e de outros componentes celulares que estão presentes (8).

Proteômica é um método direto para identificar, quantificar e estudar as modificações das proteínas de uma célula, tecido ou organismo.

O termo proteômica foi traduzido em 1995 para descrever todas as proteínas que são as expressas em um genoma (8). Definir todos os aspectos é difícil, pois o termo é mais um conceito que uma ciência bem definida atualmente.

Proteômica pode ser vista como metodologia de seleção da biologia molecular, a qual tem por objetivo documentar a distribuição geral de proteínas de célula, identificar e caracterizar proteínas individuais de interesse e principalmente elucidar as suas associações e funções. Sendo assim a proteômica fundamenta-se em princípios da bioquímica, biofísica e de bioinformática para quantificar e identificar proteínas expressas, pois elas se alteram conforme o desenvolvimento de um organismo, e também em resposta aos fatores de um ambiente (8).

A Proteômica pode ser aplicada em diversas áreas de interesse como por exemplo, na investigação de marcadores moleculares em determinadas doenças indicando a resposta de célula ou tecido a estresses externos. Através da proteômica pode-se fazer uma comparação do perfil protéico de uma célula cancerosa com o de uma célula sadia ou de um portador que está sob tratamento médico(13).

1.3 Objetivos

O objetivo deste trabalho é propor um método de classificação de câncer de ovário utilizando padrões proteômicos para auxiliar no diagnóstico junto com outros métodos como o marcador tumoral CA 125 e a ultra-sonografia na fase I, na fase inicial principalmente, pois é mais garantida a cura da neoplasia.

Propomos fazer a classificação entre tumores malignos, benignos e ausência de tumor, no caso, o grupo controle. Serão utilizadas técnicas de PCA, análise de componentes principais, para redução da dimensionalidade, e ICA, análise de componentes independentes, para extração das características. Em seguida estas características serão usadas como variáveis de entrada para um classificador baseado em redes neurais que efetuará a classificação final.

1.4 Organização do trabalho

Este trabalho está organizado da seguinte forma:

No capítulo 2 são descritos alguns conceitos e fundamentos teóricos utilizados neste trabalho, tais como processamento de marcadores através da proteômica, análise de componentes independentes, e redes neurais.

Os métodos usados neste trabalho serão mostrados no capítulo 3. No capítulo 4 os materiais.

No capítulo 5 mostraremos os resultados e fizemos algumas discussões.

Para finalizar no capítulo 6 constará as conclusões e propostas de trabalhos futuros.

Capítulo 2

Fundamentos teóricos

2.1 Os padrões proteômicos

Os padrões proteômicos surgiram com avanços na área de análise protéica que antes era realizada com técnicas como a eletroforese e que a partir 1997 tiveram um grande crescimento e com a definição da proteômica (14).

2.1.1 A Eletroforese

A eletroforese é uma técnica de separação baseada na migração das moléculas carregadas, numa solução, em função da aplicação de um campo elétrico. A eletroforese de proteínas foi realizada pela primeira vez em 1937 por Arne Tiselius, que idealizou um método denominado eletroforese livre, o qual consistia na decomposição do soro sangüíneo em cinco frações protéicas principais, trabalho que lhe rendeu um prêmio Nobel. Logo após outros métodos melhorados como eletroforese zonal e Bidimensional foram desenvolvidos(14).

2.1.2 Espectrometria de massa

É um método usado para determinar de forma precisa as massas molares. Há várias décadas, esse método vem se consolidando como ferramenta para determinação de estruturas químicas.

Um espectrômetro de massa é formado basicamente de duas partes: O sistema de ionização das moléculas, responsável por vaporizá-las e carregá-las eletricamente, e o analisador de massa. O espectrômetro de massa é que separa os íons resultantes de acordo com a massa.

Atualmente existem várias técnicas de ionização, as mais conhecidas que se complementam e se sobrepõem, dominam a análise de proteínas: a dessorção a

laser e a eletropulverização, e as técnicas mais comuns de analisador de massa são as técnicas conhecidas como *time-of-flight* (tempo de vôo), quadripolo e aprisionamentos de íons.

O método utilizado para aquisição da base de dados foi o SELDI – TOF e este utiliza a dessorção a laser para o sistema de ionização e utiliza a técnica *time-of-flight* (tempo de vôo) como analisador de massa.

2.1.2.1 SELDI -TOF

O SELDI –TOF (do inglês surface-enhanced laser desorbtion and ionization), a melhor tradução seria superfície melhorada para desorsão e ionização a laser. Este tipo de ferramenta analítica proteômica é uma classe de instrumento de espectrometria de massa. Esta técnica utiliza um chip onde são depositadas as amostras de proteínas juntamente com um ácido. Independente das manipulações iniciais, apenas um subconjunto das proteínas na amostra se ligam à superfície do chip.

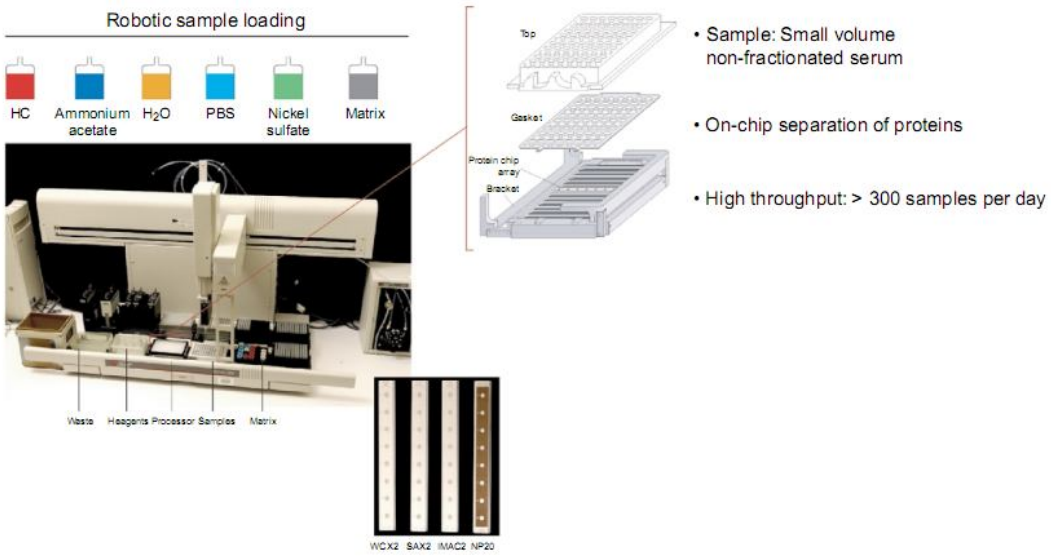

Figura 3: Espectrômetro de massa utilizado para obtenção de padrões proteômicos. Fonte: adaptada de (15).

No procedimento o chip contendo múltiplas amostras, é inserido na câmara de vácuo onde cada ponto de amostra é irradiado com um laser. O laser é irradiado

nas moléculas que estão no ácido, e sua energia causa a dessorção da molécula . Na dessorção o acído transfere prótons para molécula de proteina e a mistura passa para o estado gasoso. É condição fundamental que as moleculas estejam no estado gasoso e ionizadas para que possam ser analizadas pelo espectrometro de massa(13)(16)(17).

Conforme é apresentado na figura 2 a ampliação contem o formato do chip utilizado para aplicação de ácido para receber as amostras.

O analisador de massa utilizado usa a técnica *time-of-flight* (tempo de vôo), sistema em que moléculas ionizadas aceleradas são lançadas em tubo sob o vácuo e sem campo elétrico para medida do seu tempo de vôo até um detector. Esse tempo de vôo é proporcional a massa molar da molécula. Esse analisador é usualmente associado à dessorção a *laser*; mas pode ser também utilizado com a eletropulverização.

2.2 Análise de componentes principais

A análise de componentes principais (PCA, do inglês principal componentes analysis) é uma técnica estatística muito eficaz que pode ser usada para analisar a correlação entre os dados, ou seja, determinar as principais direções dos mesmos. Entende-se que as direções principais são vetores ortogonais sobre os quais os dados apresentam maior variância. O primeiro vetor apresenta a direção de máxima variância, o segundo vetor também está disposto conforme a direção de máxima variância, mas sob a condição de ser ortonormal ao primeiro, e assim sucessivamente ao restante dos vetores.

Uma das principais aplicações da PCA é a redução de dimensionalidade através da eliminação das variáveis originais de menor variância ou mudança do espaço da base de dados. Embora a variabilidade total de um sistema seja definida por *n* variáveis, geralmente muito desta variabilidade pode ser explicada por um número bem menor, k, de componentes principais. Desta forma a quantidade de informação contida em k é equivalente à existente nas *n* variáveis originais.

Isso nos leva a concluir que em muitas aplicações a PCA é utilizada como uma espécie de pré- processamento dos dados, servindo como entrada para outros

modelos numéricos, tais como redes neurais e outros como, análise discriminante e máquinas de vetor de suporte e lógica fuzzy. A vantagem em nosso trabalho é a redução da dimensão dos dados causada pela redução do número de parâmetros do modelo após o processamento deste pela PCA, tornando assim menor a quantidade de parâmetros da saída e melhorando o desempenho e poupando tempo de processamento para os próximos algoritmos.

2.2.1 Cálculo das componentes principais

Tomando-se matrizes e vetores como letras maiúsculas e minúsculas respectivamente, ambos em negrito. Então, matematicamente, consideremos a combinação linear:

$$z = V^T x \tag{1}$$

A matriz de autocovariância de z deve, pois, ser diagonal, de forma que:

$$C_z = E\{Z.Z^T\} = \Lambda \tag{2}$$

Sabe-se , porém, que:

$$C_z = E\{Z.Z^T\} = E\{V^T.z.z.V\} = V^T E\{x.x^T\}.V = V^T C_x V \tag{3}$$

Das equações 2 e 3, tem-se que:

$$V^T C_x V = \Lambda \tag{4}$$

Portanto, $\boldsymbol{V}^T$é a matriz ortogonal com nxk elementos que diagnaliza a matriz Cx. Como resulatdo clássico da álgebra, $\boldsymbol{V}^T$ é a matriz cujas linhas são os autovetores da matriz de Cx, correspondentes aos autovalores em ordem crescente de variância. V é uma matriz diagonal nxn cujos elementos são os autovalores de variância Cx ou seja, as variâncias de z em ordem decrescente de energia.

2.3 Análise de componentes independentes

A Análise de Componentes Independentes (Independent Component Analysis-ICA) é uma técnica que é vista como uma extensão da Análise de

Componentes Principais (Principal Component Analysis-PCA). A ICA foi desenvolvida no contexto de separação cega de fontes (Blind Source Separation-BSS), em que o problema é definido na estimação da saída de uma fonte conhecida, quando esta fonte recebe vários sinais desconhecidos e misturados. ICA tem sido aplicada em diversas áreas, como por exemplo: áudio, radar, instrumentação, médica, comunicação móvel, engenharia biomédica, e outras.

ICA é utilizada em BSS porque consegue recuperar as fontes "não-observáveis" de uma mistura de diversas fontes. O termo *blind* refere-se ao fato de que existem fontes não observáveis no sinal e nada ou pouca informação se tem sobre a mesma. Uma aplicação interessante de BSS é o problema *cocktail party*, em que separam-se as fontes originais de um sinal misturado, sem o conhecimento prévio dos coeficientes de mistura, nem a provável distribuição do sinal, usando apenas independência estatística como critério de separação de fontes (18).

Após o desenvolvimento o primeiro algoritmo de aprendizado para BSS (19). Em 1995 desenvolveram uma rede neural capaz de aprender regras que minimizam a informação mútua dos não de saída(20). Em (21) é proposto algumas variações não-lineares de PCA, e demonstraram a utilidade destes algoritmos para estimação de frequência sinusoidais. *Blind Source Separation* apresenta um grande problema na engenharia, pois a técnica mais utilizada anteriormente era PCA, que utiliza apenas estatística de segunda ordem, o suficiente apenas para descorrelacionar um conjunto de dados, mas não é necessário para independência, que requer estatística de alta ordem. Por esta razão, a ICA é vista como um método mais "robusto" que PCA, pois se PCA consegue descorrelacionar as fontes não observáveis, ICA consegue deixá-los mútua e estatisticamente independentes entre si.

2.3.1 Definições

Atribuindo que sejam observadas n misturas lineares $x_1, \dots, x_n$ de n componentes independentes.

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \dots + a_{jn}s_n \qquad j = 1, \dots, n \tag{5}$$

e que cada mistura x_j, assim como cada componente independente s_j, seja uma variável aleatória e a_j os coeficientes (pesos) da mistura linear.

Assume-se que tanto as variáveis da mistura quanto aquelas das componentes independentes têm média zero. Por conveniência, será usada a notação vetorial em vez de somas, como aquelas vistas na equação 5, utilizando letras minúsculas e maiúsculas, para representas, respectivamente, vetores e matrizes. Dessa maneira, podemos reescrever a equação 6 da seguinte maneira:

$$x = \boldsymbol{A}s \tag{6}$$

A finalidade da técnica é recuperar as fontes s, através de x, sem nenhuma informação sobre as propriedades de $\boldsymbol{A}$.

O modelo estatístico definido na equação 6 é chamado de modelo de análise de componentes independentes. Esse modelo descreve dados observados pelo processo de mistura das componentes independentes s_i, que não podem ser observadas diretamente. É preciso estimar tanto s quanto a matriz de mistura $\boldsymbol{A}$, que também é desconhecida, pois tudo que se observa é o vetor de aleatório x.

Vemos que o problema do modelo de dados de ICA, é estimar a matriz $\boldsymbol{A}$ usando apenas a informação contida na matriz x. Para tanto, é preciso fazer suposições tão gerais quanto possível (22) Portanto supõe-se que:

a) As componentes s_i são estatisticamente independentes;
b) As componentes têm distribuições não-gaussianas;
c) Torna-se mais simples usar a matriz $\boldsymbol{A}$ seja quadrada.

2.3.2 Definição de independência

Considerando y_1 e y_2 duas variáveis aleatórias. Tais variáveis são ditas independentes se a ocorrência de y_1 não influenciar na ocorrência ou não ocorrência de y_2, e vice-versa. Independência estatística é definida em termos de densidade de probabilidade. Seja $p(y_1, y_2)$ a função densidade de probabilidade (pdf) conjunta de y_1 e $\boldsymbol{s_i}$. Então, $p_1(y_1)$ denota a pdf de y_1 :

$$p_1(y_1) = \int p(y_1, \mathbf{y_2}) d_{y^2} \tag{7}$$

E similarmente para $\mathbf{y_2}$. Duas variáveis são estatísticamente independemtes se e somente se a pdf conjunta for:

$$p(y_1, y_2) = p_1(y_1) p_2(y_2) \tag{8}$$

A definição de descorrelação de duas variáveis aleatórias y_1 e $\mathbf{y_2}$, com covariância zero, é expressa por:

$$E(y_1, y_2) = E(y_1) E(y_2) = 0 \tag{9}$$

Podemos considerar que se duas variáveis são independentes, também são descorrelacionadas, mas o contrário não é válido.

2.3.3 Técnicas de estimação das componentes

Para exemplificar tomamos n misturas lineares $x_1, \dots, x_n$ de n componentes. Para estimar as componentes deve-se encontrar a inversa da matriz $\boldsymbol{A}$, que é chamada W. A solução para estimar as componentes independentes pode ser descrita da seguinte forma:

$$y = Wx = W\boldsymbol{A}s \rightarrow DPs \tag{10}$$

Sendo P uma matriz de permutação qualquer e D uma matriz diagonal não singular. Observa-se imediatamente que $\boldsymbol{y_i}$ é uma combinação linear s_i. Baseado no teorema do limite central $\boldsymbol{y_i}$ é mais gaussiano que s_i e torna-se menos gaussiano, quando de fato é igual a uma componente de s.

Para estimar as componentes independentes, deve-se finalmente encontrar a matriz W que minimiza a não gaussianidade de Wx.

Um elemento chave para a estimação do modelo de ICA é a não-gaussianidade, pois a matriz $\boldsymbol{A}$ não é identificável quando as componentes

independentes têm distribuição gaussiana. Se considerarmos que o vetor x é distribuído de acordo como o modelo de ICA na equação 6, e que todas as componentes independentes têm distribuições iguais. Para estimar as componentes, basta encontrar as combinações lineares corretas das variáveis da mistura x_i, de modo que:

$$s = A^{-1}x \tag{11}$$

Dessa forma, pode-se expressar uma combinação linear de x_i por:

$$y = b^T x \tag{12}$$

$$= \sum_i b_i \, x_i \tag{13}$$

$$= b^T \boldsymbol{A} s \tag{14}$$

Em que b deve ser determinado. A partir da equação 17 pode-se observar que y é uma combinação linear de s_i , com coeficientes dados por $q = b^T \boldsymbol{A}$. Obtêm-se:

$$y = q^T s \tag{15}$$

$$= \sum_i q_i \, s_i \tag{16}$$

Caso b corresponda a uma das linhas da inversa de $\boldsymbol{A}$, então y será uma das componentes e nesse caso, apenas um dos elementos de q será igual a 1. Enquanto todos os outros serão iguais a zero. Não é possível determinar b exatamente, mas pode-se estimar seu valor com boa aproximação.

Uma maneira de determinar b é variar os coeficientes em q e então verificar como a distribuição de $y = q^T s$ muda. Já que, conforme o Teorema do limite central (23), a soma de duas variáveis aleatórias independentes é mais gaussiana que as variáveis originais, $y = q^T s$ normalmente é mais gaussiana que qualquer uma das s_i e menos gaussiana quando se iguala a uma das s_i. Nesse caso, apenas um dos elementos q_i de q é diferente de zero (22).

Os valores de q são desconhecidos na prática e sabe-se que, através das equações 12 e 15, que:

$$b^T x = q^T s \tag{17}$$

podemos variar b e observar a distribuição de $b^T x$. Portanto, pode-se usar, como b, um vetor que maximiza a não-gaussianidade de $b^T x$, sendo que esse valor necessariamente corresponde a $q = \boldsymbol{A}^T s$, vetor esse que possui apenas uma das componentes diferente de zero. Isso significa que y na equação 12 é igual a uma das componentes independentes. Por isso, a maximização de $b^T x$ permite encontrar uma das componentes.

2.3.4 Negentropia como medida de não-gaussianidade

Uma medida importante de não-gaussianidade é a negentropia, que é baseda na entropia. Tomando um vetor aleatório y cuja função densidade de probabilidade é $f(y)$, tem–se a entropia dada por:

$$H(y) = -\int f(y) \log f(y) \tag{18}$$

A variável gaussiana tem maior entropia entre todas as variáveis aleatórias de igual variância (23). De acordo com os resultados fundamentais da teoria da informação. Isso quer dizer que uma versão modificada da entropia diferencial pode ser usada como medida de não gaussianidade. Essa medida é chamada negentropia, definida por :

$$J(y) = H(y_{gauss}) - H(y) \tag{19}$$

em que y_{gauss} é uma variável aleatória de mesma matriz de covariância de y. A negentropia é sempre não-negativa, tem valor igual a zero, se e somente se y tem distribuição gaussiana e é invariante para transformações lineares inversíveis.

Em contradição às suas qualidades como medida de não-gaussianidade, a negentropia é de difícil estimação. Por isso, é necessária a utilização de aproximações usando, por exemplo, momentos de alta ordem. Logo:

$$J(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} kurt(y)^2 \tag{20}$$

Sendo $kurt(y)$, a curtose de y, é definida como momento de quarta ordem da varável aletória y, expresso por:

$$kurt(y) \approx E\{y^4\} - 3\,(E\{y^2\})^2 \tag{21}$$

Entretanto, essa aproximação usa a curtose, que é uma medida de não-gaussianidade. Dessa forma, é mais conveniente utilizar outras abordagens, que inclusive substituem os momentos polinomiais y_3 e y_4 por outra função G . O método propõe a aproximação da negentropia baseado em expectâncias E (22).

$$J(y) \approx k_1(E\{G_1(y)\})^2 + k_2(E\{G_2(y)\} - E\{G_2(v)\})^2 \quad (22)$$

Sendo k_1 e k_2 contantes positivas, v é uma variável gausiana de média zero, e $G_1(y) = y^3$e $G_2(y) = y^4$.

2.4 Seleção de Características mais significantes

Para encontrar a melhor combinação de características (variáveis) foi utilizado o método foward-selection, iniciando com uma única característica, e incrementando mais características, passo a passo (24). Dessa forma, cada característica é adicionada no modelo de cada vez. A cada passo, cada característica que ainda não pertence ao modelo é testada para ser incluída. As características mais signifcativas, ou seja, que mostram um maior decréscimo na função de erro descrita acima são adicionados ao modelo, até que se consiga um subconjunto p , menor que o conjunto P selecionado.

Na técnica Forward-Selection, cada etapa envolve o crescimento da rede pelo acréscimo de uma função base (ou seja, uma característica). Adicionar uma função base nova é uma das operações incrementais. A equação fundamental da técnica é:

$$P_{m+1} = P_m - \frac{P_m . f_j . f_j^T P_m}{f_j^T . P_m f_j} \quad (23)$$

que expressa a relação entre a relação entre P_m , a matriz de projeção de m camadas escondidas do subconjunto corrente e P_{m+1} , a projeção sucedente do membro do último conjunto de características adicionado. Os vetores $\{f_j\}_{J=1}^{M}$ são colunas de um conjunto selecionado de funções-bases

$$F = [f_1, f_2, f_M] \quad (24)$$

sendo $M \gg m$.

A escolha das funções bases é baseada em encontrar erro-médio-quadrático. Das regras de atualização para a matriz de projeção e da equação para o erro:

$$T_m - T_{m+1} = \frac{(y^T P_m \cdot f_j)^2}{f_j^T \cdot P_m f_j} \tag{25}$$

sendo T_m o conjunto anterior e T_{m+1} o conjunto atual.

2.5 Redes neurais artificiais com classificadores

Na década de 40 as redes neurais artifciais (RNAs) foram desenvolvidas, originalmente, pelo neurofisiologista Warren McCulloch e pelo matemático Walter Pitts uma analogia entre células nervosas vivas e o processo eletrônico num trabalho publicado sobre "neurônios formais", que consistia num modelo de resistores variáveis e amplificadores representando conexões sinápticas de um neurônio biológico.

A partir da década 80, outros modelos de redes neurais artificiais foram elaborados com o propósito de aperfeiçoar e aplicar esta tecnologia. A Figura 4 apresenta a arquitetura de uma RNA.

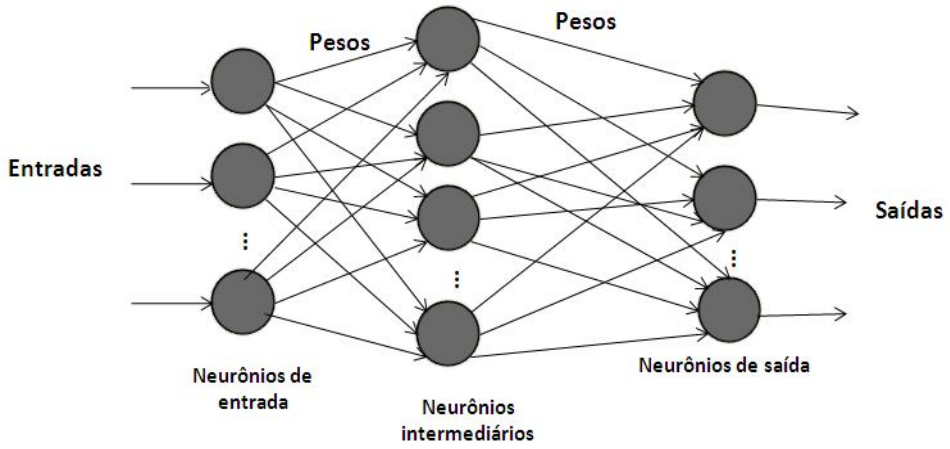

Figura 4: Arquitetura de uma Rede neural artificial RNA

Existem muitas variantes de uma rede neural e combinando-as, pode-se mudar a arquitetura conforme a necessidade da aplicação. Os itens que compõem uma rede neural de forma geral, portanto, sujeito a modificações, são os seguintes:

- Conexões entre camadas;

- Camadas intermediárias;

- Quantidade de neurônios;

- Função de transferência;

- Algoritmo de aprendizado;

Uma RNA deve possuir alguma regra de treinamento, em que os pesos de suas conexões são ajustados de acordo com os padrões apresentados, de tal forma que pode-se dizer que uma importante propriedade é a capacidade do aprendizado. Isso é feito através de um processo iterativo de ajustes aplicados aos pesos das conexões ao qual denomina-se treinamento. O aprendizado ocorre quando a RNA atinge uma solução generalizada para uma classe de problemas.

Para que uma RNA tenha a capacidade de aprender esta deve possuir um algoritmo de aprendizado que tenha regras bem definidas. Existem muitos algoritmos de aprendizado específicos para determinados modelos de redes neurais, estes algoritmos diferem entre si principalmente pelo modo como os pesos são modificados. Outro ponto importante é a categorização das situações de aprendizado das RNAs. Neste contexto pode-se citar as seguintes classes de aprendizado (25)(26):

- Aprendizado supervisionado, quando é utilizado um agente externo que indica à rede a resposta desejada para o padrão de entrada;

- Aprendizado não supervisionado, quando não existe um agente externo indicando a resposta desejada para os padrões de entrada.

2.6 Redes neurais Perceptron multicamadas

As redes neurais Perceptron multicamadas (MLP) são as redes neurais mais frequentemente usadas em reconhecimento de padrões (27), (28). A estrutura de

uma RNA do tipo MLP é constituída por um conjunto de nós fonte, os quais formam a camada da entrada da rede (*input layer*), uma ou mais camadas escondidas (*hidden layers*), e uma camada de saída (*output layer*), que extraem informações durante o aprendizado, e atribuindo coeficientes ou pesos às camadas de entrada.

O número de nós fonte na entrada da rede é determinado pela dimensionalidade do espaço de observação, que é responsável pela geração dos sinais de entrada. O número de neurônios na camada de saída é determinado pela dimensionalidade requerida na resposta desejada. Assim, o projeto de uma rede MLP requer a consideração de três fatores:

a) A determinação do número de camadas escondidas;

b) A determinação do número de neurônios em cada uma das camadas escondidas;

c) A especificações dos pesos sinápticos que interconectam os neurônios nas diferentes camadas da rede.

Os fatores a e b determinam a complexidade do modelo da RNA escolhida, e infelizmente, não há regras determinadas para tal especificação. A função das camadas escondidas em uma RNA é a de incluir na relação entrada-saída da rede de uma forma ampla. Uma RNA com uma ou mais camadas escondidas é apta a extrair as características de ordem superior de algum desconhecido processo aleatório subjacente, responsável pelo comportamento dos dados de entrada, processo sobre o qual a rede está tentando adquirir conhecimento. A RNA adquire uma perspectiva global do processo aleatório, apesar de sua conectividade local, em virtude do conjunto adicional de pesos sinápticos e da dimensão adicional de do número de interações proporcionada pelas camadas escondidas. O fator c envolve a utilização de algoritmos de treinamento supervisionado. As rede neurais artificiais MLPs têm sido aplicadas na solução de diversos e difíceis problemas através da utilização de tais algoritmos. O algoritmo de treino geralmente utilizado é o algoritmo de retropropagação, popularmente conhecido como backpropagation. O algoritmo backpropagation baseia-se na heurística do aprendizado por correção de erro. Este algoritmo pode ser visto como uma generalização do algoritmo LMS (Least Mean Square), desenvolvido por Bernard Windrow (29). Basicamente, o algoritmo

backpropagation consiste de dois passos através das diferentes camadas do MLP: um passo direto e um passo reverso.

No passo direto, um padrão de atividade do processo a ser aprendido (ou vetor de entrada) é aplicado ao nós de entrada do MLP e o seu efeito se propaga através da rede, camada por camada, produzindo na camada de saída a resposta a excitação aplicada (vetor de saída). Durante o passo direto os pesos sinápticos são todos fixos. No passo reverso, os pesos sinápticos são todos ajustados de acordo com a regra de aprendizado por correção de erro. Ou seja, a resposta do MLP à excitação é subtraída de um padrão de resposta desejado para aquela excitação aplicada, de forma a produzir um sinal de erro, de forma semelhante ao do algoritmo LMS. Este sinal de erro é, então, propagado de volta aos mesmos neurônios utilizados no passo direto, porém no caminho contrário do fluxo de sinal nas conexões sinápticas, daí o nome *backpropagation*. Os pesos sinápticos são, então, ajustados de forma que a resposta obtida da MLP aproxime-se mais do padrão de resposta desejado.

Durante o treinamento as MLPs constroem um espaço multidimensional definido pela ativação dos nós das camadas escondidas, de modo que as classes sejam mais separáveis possível. O modelo da superfície de separação se adapta aos dados.

2.7 Redes neurais de função de base radial

As redes RBF (função de base radial) têm sido tradicionalmente usadas em redes neurais com uma única camada intermediária, como a mostrada na figura 5.

Cada um dos n componentes do vetor de entrada x são aplicados às m funções de ativação, cujas saídas são linearmente combinadas com pesos w_j, com j variando de 1 a m. Para o caso de uma única saída $y = f(x)$, o mapeamento entrada e saída é dado por:

$$y = \sum_{j=1}^{m} w_j h_j\left(\vec{x}\right) \tag{26}$$

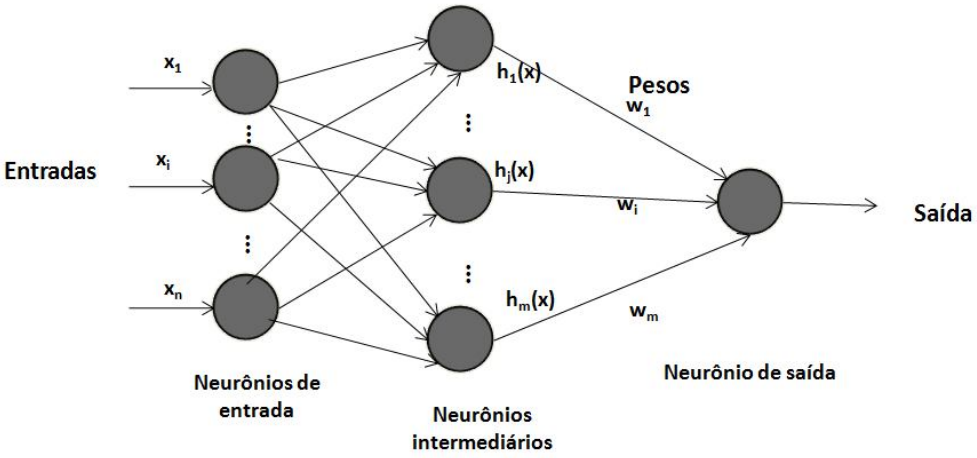

Figura 5: Arquitetura da rede neural artificial com função de ativação de base radial (RBF).

Se os centros e as aberturas das funções de ativação forem fixos e apenas os pesos da camada de saída forem ajustáveis, o modelo será linear nos parâmetros. Caso os centros e as aberturas sejam ajustáveis, o modelo será não-linear nos parâmetros.

As funções de base radial (RBF) são uma classe especial de funções. São caracterizadas por uma resposta que decresce ou cresce monotonicamente com a distância a um ponto central. Os parâmetros mais importantes a serem definidos em uma função de base radial são o centro e a taxa de crescimento, ou decrescimento, da função(30).

Uma função de base radial típica é a Gaussiana. Para o caso de uma entrada escalar, a Gaussiana é dada pela expressão:

$$h(x) = \frac{\sqrt{r^2 + (x-c)^2}}{r} \tag{27}$$

Nas funções mostradas acima, o parâmetro c corresponde ao centro e o parâmetro r é uma medida de abertura.

Capítulo 3

Método

3.1 Metodologia

Todas as técnicas descritas anteriormente foram aplicadas no método desenvolvido. A partir dos padrões proteômicos de cada paciente serão extraídas as características, estas são retiradas utilizando ICA, que também faz uma redução da dimensionalidade da amostra utilizando PCA que já está incorporado no ICA, daí então a matriz de amostras reduzida, ou seja, as características alimentam redes neurais, em que a seleção de características mais significantes será feita utilizando a técnica Forward-Selection e a decisão final do diagnóstico, trabalhamos usando redes neurais Multilayer Perceptron e a metodologia é mostrada na figura 6.

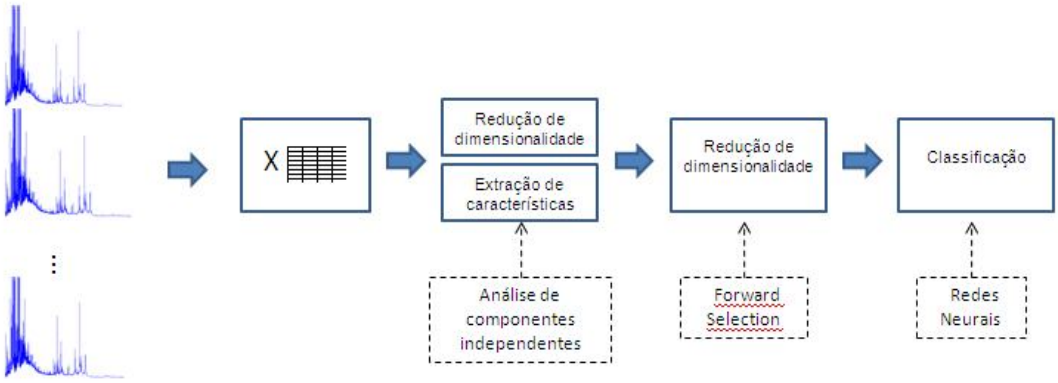

Figura 6: Etapas da metodologia de trabalho

3.1.1 Aquisição de dados

Os dados são baseados em padrões proteômicos usando a técnica SELDI-TOF que por sua vez é uma padrão de informação mais preciso para auxiliar no

diagnóstico, é uma estrutura de dados baseada no peso molecular em função do tempo gasto de locomoção destas moléculas analisadas pelo espectrômetro e a amostra possui 15.154 pontos o que provoca uma taxa de processamento alta se estes dados forem usados diretamente, sem um pré-processamento.

3.1.2 Redução da dimensionalidade

Neste trabalho cada amostra de pacientes é disposta de 15.154 pontos o que gera para cada pacientes um vetor grande. É importante utilizar uma técnica para reduzir a dimensionalidade dos dados, utilizou-se PCA que já está incluso no próprio ICA. Os experimento foram simulados com duas bases de dados, DB1 e DB2, gerando matrizes de 100 x100 e 91x91, respectivamente, com o objetivo de facilitar a extração de características e classificação.

3.1.3 Extração de Características

Após a redução da dimensão que já está incorporada no ICA, o ICA segue com projeção de novas funções bases, utilizaremos estas funções para estimar os coeficientes, e estas serão tratadas como características do grupo de câncer ou controle.

3.1.4 Classificação

Como último passo, serão usados dois modelos de redes neurais como classificador: Rede neural artificial multicamadas Perceptron e rede neural artificial com função de ativação base radial para classificar os amostras dos pacientes em controle e câncer. Ambos os classificadores foram descritos no capítulo anterior.

Para fins de comparação, as redes neurais receberam o mesmo conjunto de treinamento e teste, escolhidos entre as amostras de forma aleatória.

3.1.5 Validação do método de Classificação

Para avaliar o classificador em relação à sua capacidade de diferenciação foi utilizado a sua sensibilidade, especificidade e acurária. Na discussão dos resultados será abordado de forma mais detalhada o conceito e forma de calcular estas medidas de validação de resultado.

Capítulo 4

Materiais

4.1 Base de dados

Foram utilizados duas bases de dados do centro de pesquisa do câncer, “Center cancer research”, do Instituto Nacional do Câncer nos EUA(31). Estas são amostras de padrões proteômicos, extraídas através da técnica SELDI-TOF utilizando um chip WCX2. Na segunda amostra houve uma diferença na manipulação no processo de lavagem e incubação interferindo na qualidade espectral, o que foi chamado de upgrade pBSII no SELDI-TOF. A primeira conta com amostras de 100 pacientes do grupo com câncer, 100 pacientes do grupo controle e 16 pacientes de câncer benigno. A segunda contém amostras de 162 pacientes do grupo com câncer e 91 pacientes do grupo controle.

4.2 Softwares utilizados

Após a seleção das amostras de pacientes dos padrões proteômicos, utilizamos o software MatLab, para gerar a matriz de coeficientes A, através do algoritmo *FastIca*.

Utilizando o *FastIca*, estamos internamente também processando o PCA, e este foi essencial para redução da dimensionalidade dos dados de entrada para o classificador.

Como classificador utilizamos o programa *Trajan Neural Networks* que utiliza um pré-processamento das entradas que é caracterizado por uma etapa de seleção de características encontradas de cada nova amostra, ou seja, cada linha da matriz

de coeficientes A, já que o vetor de entradas para rede neural ficaria grande, o que poderia comprometer o desempenho da rede. A seleção de características foi realizada através do algoritmo *Foward-Selection.* Depois de escolhidas as características mais significantes, estas serviram de entrada para a rede neural artificial Perceptron multicamadas e para rede neural com função de ativação de base radial, para que fosse dada a decisão final de diagnóstico.

Capítulo 5

Resultados e Discussão

Neste capítulo, será descrito como foram aplicadas as técnicas no método proposto, os resultados encontrados e a discussão destes.

5.1 Variáveis selecionadas

Como já foi mencionado no item 3.1 trabalhou-se com duas base de dados: a primeira conta com amostras de 100 pacientes do grupo com câncer, 100 pacientes do grupo controle e 16 pacientes de câncer benigno, definiu-se por DB1, a segunda contém amostras de 162 pacientes do grupo com câncer e 91 pacientes do grupo controle chama-se por DB2. Para facilitar selecionou-se 100 pacientes do grupo câncer e 100 pacientes do grupo controle, o que definiu-se SDB1 e para DB2 utilizou-se 91 amostras de pacientes do grupo câncer e 91 do grupo controle retirando as demais, define-se a segunda por SDB2.

Ressalta-se ainda que cada amostra de pacientes é disposta de 15.154 pontos o que gera para cada pacientes um vetor diferente de 1X15.154.Existe ainda uma escala de variabilidade do espectrômetro de massa mas como esta é idêntica a todas as amostras foi retirada do processamento visto que esta apenas servirá como base para transformação na grandeza de peso molecular.

Neste item será mostrado o passo a passo do método de extração de características, reforça-se ainda que este método possui o PCA incluso o que resulta na redução da dimensionalidade dos vetores de amostra de cada paciente que 1x 15154 passa para 1x100 ou 1x91.

5.2 Extração de parâmetros usando ICA

Consideramos que cada amostra é estatisticamente independente e que cada amostra seja definida pela função :

$$x_{i=.}a_{i1}.s_1 + a_{i2}.s_2 + \cdots + a_{in}.s_n \tag{28}$$

No modelo acima apenas as variáveis x_i são conhecidas, e a partir delas serão estimados os coeficientes da mistura a_i e as componentes s_j, ou seja :

$$x = \boldsymbol{A}s \tag{29}$$

Sendo x , a matriz de mistura, $\boldsymbol{A}$ os coeficientes e s as componentes independentes que compõem a amostra.

Na equação x representada a função base de cada paciente.

Cada amostra representa uma linha da matriz de mistura. A matriz x é representada por amostras na dimensão de P, ou seja, 1x50. Então, cada linha da matriz A corresponde a uma amostra e cada coluna corresponde a um peso atribuído para a amostra, ou seja, um parâmetro de entrada para a rede neural.

O algoritmo utilizado para fazer a extração de parâmetros através de ICA foi o fastICA, que um algoritmo comumente utilizado para fazer separação de fontes cegas (BSS), onde estimamos as funções bases a partir da matriz de mistura X.

Este algoritmo é baseado em interações de ponto fixo(32)(33)(34), o algoritmo do ponto fixo foi introduzido usando curtose(34)(33), o algoritmo *FastICA* foi generalizado para outras funções. Para dados pré-processados por branqueamento (whitering), o algoritmo *fastICA* tem a seguinte forma:

$$w(K) = E\{x.g(w(k-1)^T x)\} - E\{g'(w(k-1)^T.x)\}w(k-1) \tag{30}$$

Em que o vetor de pesos w é então normalizado para a norma unitária após cada interação, e

$$w^{-1} = \boldsymbol{A} \tag{31}$$

A função g é derivada da função G usada nas funções de custo em geral ou seja

$$J_{G(y)} = \left|E_y\{G(v)\}\right|^p \tag{32}$$

Sendo v uma variável aleatória, y normalizado para variância unitária e o expoente p = 1,2.

Uma forma básica do algoritmo ICA é dado como segue:

1. Escolha um vetor inicial aleatório w.
2. Faça $w^+ = E\{xg'(w^Tx)\} - E\{g'(w^Tx)\}w$
3. Faça $w = \frac{w^+}{\|w^+\|}$
4. Se não convergir, volta ao passo 2.

As esperanças estimadas, na prática a média das amostras sobre a quantidade suficiente dos dados de entrada.

Dentre as vantagens do algoritmo FastICa, podemos citar a rápida convergência, a simplicidade do algoritmo, e seu processamento paralelo, distribuído, computacionalemente simples, que requer pouco espaço de memória.

Usando o FastICA e a matriz x, obtemos a matriz de funções bases a, que contém as características de cada amostra.

Figura 7: Sinal proteômico de 3 pacientes, plotagem do espaço bidimensional.

5.3 Configuração da rede neural

O programa Trajan Neural Networks selecionou as duas melhores configurações de redes neurais MLP. O número de neurônios na camada escondida foi incrementado até 50, porém a rede com 21 neurônios apresentou melhor desempenho. O erro encontrado no treinamento foi de 0,008, que, segundo o programa utilizado, indicou que a rede neural escolhida já estava especializada. Sendo assim, a melhor rede neural MLP encontrada ficou com a seguinte configuração: 45 neurônios na camada de entrada, 21 na camada escondida e 2 na camada de saída, conforme é apresentado na Figura 8:

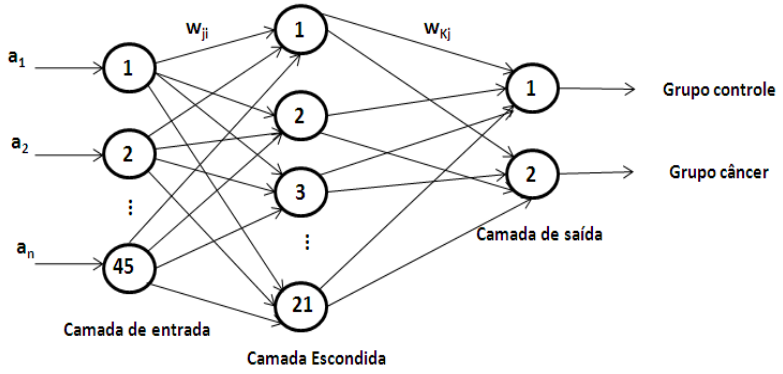

Figura 8: Configuração de uma rede neural muti-camadas Perceptron.

5.4 Configuração da rede de função base radial

O programa Trajan Neural Networks trabalha com várias configurações de redes neurais FBR. O erro de teste médio foi melhor 0,402 na com 30 neurônios. No entanto na rede com 34 neurônios o erro médio encontrado no treinamento foi menor de 0,01 que a com 30 neurônios e esta apresentou melhor desempenho no problema de classificação, segue a configuração da rede FBR com melhor desempenho: 23 na camada e entrada, 34 camada escondida e 2 na camada de saída, conforme a Figura 9:

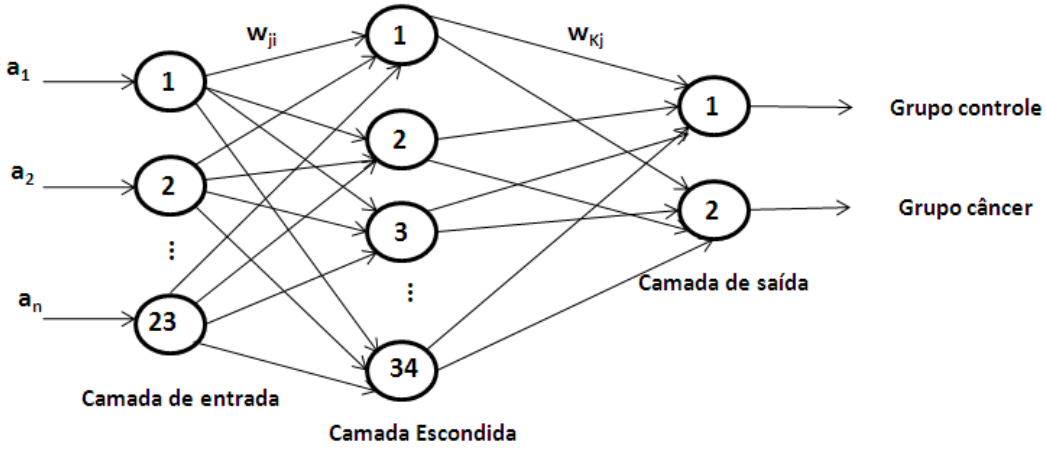

Figura 9: Configuração de uma rede neural FBR.

5.5 Testes

Foram feitos vários testes para verificação da eficácia dos resultados as tabelas 1 a 8 contém a especificidade, sensibilidade e acurária baseada nos resultados de classificação apresentado nas simulações com o programa Trajan Neural Networks que em média trabalha com duas configurações de redes neurais FBR e em média duas ou três configurações de rede MLP, geralmente modifica aleatoriamente a quantidade de neurônios na camada escondida.

As tabelas 1 a 4 mostram alguns dos testes para a primeira base de dados SDB1**.**

Tabela 1. Tabela de confusão para simulação com SDB1 – teste 1

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
PNN	94	97	3	6	97%	94%	95%
Linear	98	97	3	2	97%	98%	97%
MLP	91	88	12	9	88%	91%	89%
RBF	91	88	12	9	88%	91%	89%
RBF	91	90	10	9	90%	91%	90%

Tabela 2. Tabela de confusão para simulação com SDB1 – teste 2

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
PNN	98	92	8	2	92%	98%	95%
Linear	90	92	8	10	92%	90%	91%
MLP	98	92	8	2	92%	98%	95%
RBF	98	92	8	2	92%	98%	95%
RBF	96	90	10	4	90%	96%	93%

Tabela 3. Tabela de confusão para simulação com SDB1 – teste 3

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
Linear	86	90	10	14	90%	86%	88%
MLP	96	98	2	4	98%	96%	97%
MLP	93	97	3	7	97%	93%	95%
RBF	96	98	2	4	98%	96%	97%
RBF	96	98	2	4	98%	96%	97%

Tabela 4. Tabela de confusão para simulação com SDB1 – teste 4

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
PNN	65	88	12	35	88%	65%	76%
Linear	70	70	30	30	70%	70%	70%
RBF	97	90	10	3	90%	97%	93%
MLP	80	98	2	20	98%	80%	89%
MLP	80	98	2	20	98%	80%	89%

Foram utilizados também pelo programa de simulações outras redes como a PNN – (Rede artificial probabilística) e rede artificial linear, no entanto focamos no

trabalho somente a nível de explicação somente RBF e MLP, cujos resultados foram melhores.

As tabelas 5 a 8 mostram alguns dos testes para a segunda base de dados SDB2.

Tabela 5. Tabela de confusão para simulação com SDB2 – teste 1

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
MLP	19	78	13	72	86%	21%	53%
MLP	80	76	15	11	84%	88%	86%
Linear	87	88	3	4	97%	96%	96%
RBF	86	88	3	5	97%	95%	96%
RBF	87	88	3	4	97%	96%	96%

Tabela 6. Tabela de confusão para simulação com SDB2 – teste 2

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
MLP	19	78	13	72	86%	21%	53%
MLP	80	76	15	11	84%	88%	86%
Linear	87	88	3	4	97%	96%	96%
RBF	86	88	3	5	97%	95%	96%
RBF	87	88	3	4	97%	96%	96%

Tabela 7. Tabela de confusão para simulação com SDB2 – teste 3

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
PNN	88	85	6	3	93%	97%	95%
MLP	45	50	41	46	55%	49%	52%
Linear	80	87	4	11	96%	88%	92%
RBF	80	87	4	11	96%	88%	92%
RBF	80	87	4	11	96%	88%	92%

Tabela 8. Tabela de confusão para simulação com SDB2 – teste 4

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
PNN	90	85	6	1	0,934%	0,989%	0,962%
MLP	45	50	41	46	0,549%	0,495%	0,522%
Linear	70	64	27	21	0,703%	0,769%	0,736%
RBF	85	87	4	6	0,956%	0,934%	0,945%
RBF	84	87	4	7	0,956%	0,923%	0,940%

Baseado na tabela 9, as linhas representam as classe do classificador da rede RBF e as colunas a quantidade de pacientes para cada classe conforme SDB1, o sucesso é o percentual de acerto.

Tabela 9. Tabela com matriz de confusão com classificadores que apresentaram melhor resultados para SBD1.

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
MLP	96	98	2	4	98%	96%	97%
RBF	96	98	2	4	98%	96%	97%
PNN	94	97	3	6	97%	94%	95%

Tabela 10. Tabela com matriz de confusão com classificadores que apresentaram melhor resultados para SBD2.

Rede Neural	VP	VN	FP	FN	Especificidade	Sensibilidade	Acurácia
PNN	90	85	6	1	93%	98%	96%
RBF	87	88	3	4	97%	96%	96%

5.6 Discussão

Para entendermos melhor o que significa sensibilidade, especificidade e acurácia, pois são as medidas mais utilizadas para descrever um sistema de classificação médica, vamos definir as variáveis que auxiliarão:

Sendo FN Falso Negativo, FP Falso Positivo, VN Verdadeiro Negativo e VP Verdadeiro Positivo.

• Verdadeiro Negativo (VN) – Diagnóstico de neoplasia classificada como benigna, ou inexistente.

• Verdadeiro Positivo (VP) - Diagnóstico de neoplasia classificada como maligna.

•Falso Negativo (FN) é a uma neoplasia maligna classificada benigna ou inexistente.

• Falso Positivo (FP) é a uma não neoplasia classificada como maligna:

• Sensibilidade (S) é a proporção de verdadeiros positivos que são corretamente identificados pelo teste, e é definida por:

$$S = VP/(VP + FN) \tag{33}$$

• Especificidade (E) é a proporção de verdadeiros negativos que são corretamente identificados no teste, e é dada por:

$$E = VN/(VN + FP) \tag{34}$$

• Acurácia (A) é a taxa de sucesso ou acerto do teste e é dada por:

$$A = (VN + FP)/(VP + TN + FP + FN) \tag{35}$$

Conforme tabela , o método obteve 96 diagnósticos verdadeiro-positivos, 98 verdadeiro-negativos, 2 falso-positivos e 4 falso negativos para MLP e RBF, e ambas apresentaram os mesmos valores para especifidade 98%, sensibilidade 96% e acurácia de 97%.

Baseado na tabela 10, o método obteve 87 diagnósticos verdadeiro-positivos, 88 verdadeiro-negativos, 3 falso-positivos e 4 falso negativos para RBF e apresentou os valores para especificidade 97%, sensibilidade 96% e acurácia de 96%.

Na tabela 10 a PNN também obteve um bom resultado, no entanto não consideramos para a definição final de taxa de acertos.

Consideramos que o nosso método apresentou como taxas de acerto com os seguintes valores para especificidade 98%, sensibilidade 96% e acurácia de 97%, ou seja os valores para simulação com SDB1.

. O erro encontrado, levando em consideração as duas classes de SDB1 (controle e maligna) foi de apenas 0,98% para o teste da rede RBF e 0,64% para

rede MLP.Já para SDB2 considerando as classes(controle e maligna) o erro foi muito menor de apenas 0,022% para o teste da rede RBF e 0,0002% para rede MLP.

.

Capítulo 5

Conclusão e trabalhos futuros

Este trabalho propõe um método de classificação de padrões proteômicos utilizando análise de componentes (ICA) e redes neurais Perceptron multicamadas e redes neurais função de ativação de base radial. O método proposto utilizou o algoritmo FastICA para extrair um conjunto de características em conjunto com a técnica Foward Selection, para selecionar as características mais significantes. O conjunto destas técnicas permitiu que o resultado encontrado fosse melhor que em métodos que utilizam de outras técnicas como algoritmos genéticos e *agent swarm*.

Considerando que os padrões proteômicos são uma grande solução para detecção precoce do câncer, pois a base de dados utilizada os pacientes se encontram no estágio I. Nosso objetivo é contribuir com método que seja mais eficaz.

O presente resultado demonstrou que as técnicas utilizadas neste trabalho conseguiram classificar muito bem padrões proteômicos das classes controle(normal), câncer maligno. Observou-se também, que com as redes neurais obteve-se uma taxa de sucesso média de 97%, apresentando baixo erro, com erros menores, pode-se obter um menor número de biopsias desnecessárias, e também um menor número de casos de câncer descobertos tardiamente.

Relacionando com trabalhos anteriores, o método proposto utilizou o algoritmo FastIca em conjunto com algoritmo de seleção de características *Foward-Selection*, ao passo que no método descrito em (35), que obteve com resultado 86% e 91,3% de acurácia em experimentos separados e (36) obteve 90% de acurácia utilizando algoritmo baseado em *agent swarm*.

A partir do realizado no presente trabalho, nos próximos passos pretende-se:

- Testar a eficácia do método com uma base de dados regional, ainda em fase de construção.

- Elaborar um programa, que seja de baixo custo computacional e simples de ser manuseado, para ser utilizado por médicos e especialistas em ultra-sonografia.

- Adaptar o método para classificação de outras estruturas, que utilizem como entrada padrões proteômicos, por exemplo: câncer de próstata e de pâncreas.

Artigos Publicados pelo autor

- **Simone C. F. Neves, Lúcio Flávio Campos, Ewaldo Santana, Ginalber Serra e Allan Kardec Barros.** Diagnosis of ovarian cancer with proteomic patterns in serum using independent component analysis and neural networks – Singapore December 2010 international conference on control, automation, robotics and vision engineering, Dubai – Emirados Árabes – WASET 2010 SINGAPORE.
- **Osevaldo S. Farias, Jorge H. M. Santos, João V. F. Neto, Sofiane Labidi , Thiago Drumond, Pinheiro Moura e Simone C. F. Neves**. A Real Time Expert Systems for Faults Identification in Rotary Railcar Dumper - 5th International Conference on Informatics in Control, Automation and Robotics – Funchal, Madeira – Portugal - ICINCO 2008.
- **Pinheiro Moura, João Viana, Marco Souza, Denis Anderson, Bruno Eduardo e Simone Neves**. Simulador de operação remota virtual de equipamentos portuários do porto de Ponta da Madeira na Vale de São Luís - 15º Congresso Internacional e Exposição Sul-Americana de Automação – São Paulo, Brasil - BRAZIL AUTOMATION ISA 2011.

REFERÊNCIAS

1. Ovário. *INCA.* [Online] Instituto nacional do câncer. [Citado em: 14 de março de 2012.] www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/ovario.

2. Types of cancer ovarian. *ACS.* [Online] American Cancer Society. [Citado em: 15 de março de 2012.] http://www.cancer.gov/cancertopics/types/ovarian.

3. Ovarian cancer detailed guide. *ACS.* [Online] [Citado em: 15 de março de 2012.] http://www.cancer.org/acs/groups/cid/documents/webcontent/003130-pdf.pdf.

4. **Dr. Ricardo Caponero.** Aspectos oncológicos no carcinoma epitelial do ovário. *MEDCENTER.* [Online] [Citado em: 10 de 01 de 2012.] http://www.medcenter.com/Medscape/content.aspx?id=26409.

5. Câncer de ovário - Medscape. *Medscape.* [Online] MEDCENTER. [Citado em: 14 de março de 2012.] http://www.medcenter.com/Medscape/content.aspx?bpid=129&id=24987.

6. Epidemoilogia. *MEDCENTER.* [Online] [Citado em: 02 de 03 de 2012.] http://www.medcenter.com/Medscape/content.aspx?bpid=129&id=24989.

7. Cancêr de ovário. *News medical.* [Online] [Citado em: 14 de Março de 2012.] http://www.news-medical.net/health/Ovarian-Cancer-%28Portuguese%29.aspx.

8. **Luciana Di Ciero, Cláudia de Mattos Bellato.** Proteoma. *Biotecnologia ciência e desenvolvimento.* São Paulo, 2002, Vol. 29.

9. Sobre o câncer de ovário. *CPO.* [Online] [Citado em: 2012 de 01 de 22.] www.centropaulistaoncologia.com.br/sobre_cancer.php.

10. Câncer de ovário. *Assunto de saúde.* [Online] 22 de 01 de 2012. http://assuntodesaude.blogspot.com.br/2010/08/cancer-de-ovario-o-cancer-de-ovario-e-o.html.

11. **FERNANDES, Luís R. Araujo, LIPPI, Umberto Gazi e BARACAT, Fausto F.** Índice de Risco de Malignidade para tumores do ovário incorporando idade, ultra-sonografia e o CA-125. São Paulo, 2003.

12. *Diversidade génetica em populções em populções de Myracrodruon urundeuva Fr All, sob diferentes tipos de pertubação antrópica.* **Viegas, Michele, Luiz, Mário e Lacerda, Cristina.** São Paulo : Dissertação, 2009.

13. **Emanuel F. Petricoin, Lance A. Liotta.** SELDI-TOF-based proteomic patter diagnostics for early detection of cancer. *ELSEVIER.* Bethesda, 2004.

14. **Thales Lima Rocha, Paulo H. A. Costa, José C. C. Magalhães, Raphael G. S. Evaristo, Érico A. R. Vasconcelos, Marise coutinho,Norma Paes, Maria Silva, Maria de Fátima G.S.** Eletroforese bidimensional e análise de proteomas. *Embrapa.* Brasília, 2005., Vol. 136.

15. **Petricoin, Emanuel e Liotta, Lance.** Proteomic approaches in cancer risk and response assessment. *ELSEVIER.* 2004.

16. **Cunha, Ricardo Bastos, Castro, Mariana de Souza e Fontes, Wagner.** Espectrometria de massa de proteínas. *Biotecnologia ciência e desenvolvimento.* 2006.

17. **Emanuel F. Petricon III, Ali M. Ardekani, Ben A Hitt, Peter J. Levine, Vicent A. Fusaro, Seth M. Steinberg, Gordon B. Mills, Charles Simone, David A. Fishman, Elise C. Kohn, Lance A. Liotta.** Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet.* Bethesda, 2002, Vol. 359.

18. **Girolami, Mark e Fype, Colin.** An extended exploratory pursuit network with linear and non-linear anti-hebbian lateral connections applied to the cocktail party problem. *ELSEVIER.* Paisley, 1997.

19. **Jutten, Christian e Herault, Jeanny.** Blind separation of sources. *ELSEVIER.* 1991.

20. **Bell, Anthony e Sejnowshi, Terrence.** An information maximization approach to blind separation and blind deconvolution. *Neural computation.* Massachusets, 1995.

21. **Karhunen, Juha e Joutsensalo, Jyrki.** Representation and separation of signals using nonlinear PCA type learning. *Neural network.* 1997.

22. *Independent components analysis.* **Aapo Hyvärinen, Oja Karhunen.** John Wiley & Sons, Nova York : s.n., 2001. 481p.

23. **Papoulis, Athanasios e Pillai, Unnikrishma.** *Random variables and stochastic processes.* Nova York : Mc Graw-Hill, 2002.

24. **Orr, Mark;.** BRF. *Radial Basis Function Networks.* [Online] [Citado em: 14 de 02 de 2012.] http://www.anc.ed.ac.uk/rbf/.

25. **Braga, Antonio de Pádua.** *Redes Neurais e Artificiais - Teoria e Aplicações.* s.l. : LTC, 2000.

26. **Krose, Ben e Smagt, Patrick.** *An introduction to neural networks.* s.l. : University Amsterdã, 1996.

27. **Duda, Richard O. e Hart, Peter E.** *Pattern classification and scene analysis.* New York : Wiley- Interscience Publication, 1973.

28. **Bishop, Christopher M.** *Neural Networks for Pattern Recognition.* New York : Oxford University Press, 1999.

29. **Windrow, B. e Stearns, Samuel D.** *Adaptive Signal Processing.* s.l. : Prentice-Hall signal processing series, 1985.

30. **Principe, José C., Euliano, Neil R. e Lefebvre, W. Curt.** *Neural e Adaptative system.* 2000.

31. PPatterns. *Center cancer research.* [Online] [Citado em: 2010 de março de 03.] http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp.

32. *A fast fixed-point algorithms for independent component analysis.* **Aapo Hyvärinen, Erkki Oja.** Filand : s.n., 1997. pages1483-1492.

33. *A family of fixed-point algorithms for independent component analysis.* **Hyvärinen, Aapo.** Filand : s.n., 1997. pages 3917-3920.

34. *Fast and robust fixed-point algorithms for independent component analysis.* **Hivärinen, Aapo.** Filand : IEEE Trans. on Neural Networks, 1999.

35. **Loo, Lit-Hsin e Quin, John.** Classification of SELDI-TOF mass spectra of ovarian cancer serum samples using a proteomic patterns recognizer. *IEEE.* Philadelphia - USA, 2003.

36. **Meng, Yan.** A swarm intelligence based algorithm for proteomic pattern detection of cancer ovarian. *IEEE.* NewJersey - USA, 2006.