RECIIS
Revista Eletrônica de Comunicação Informação & Inovação em Saúde
[www.reciis.cict.fiocruz.br]
ISSN 1981-6278

SUPLEMENTO – BIOINFORMÁTICA E SAÚDE

Artigos originais

Máquina de agrupamento por elipsóide: uma linha de frente para auxiliar no diagnóstico de doenças

DOI: 10.3395/reciis.v1i2.Sup.101pt

Paulo Costa Carvalho
Programa de engenharia de sistemas e computação, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
carvalhopc@cos.ufrj.br

Juliana de Saldanha da Gama Fischer
Instituto de química da Universidade Federal do Rio de Janeiro e Rede Proteomica do Rio de Janeiro, Rio de Janeiro, Brasil
juli_f@iq.ufrj.br

Valmir C. Barbosa
Programa de engenharia de sistemas e computação, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
valmir@cos.ufrj.br

Gilberto Barbosa Domont
Instituto de química da Universidade Federal do Rio de Janeiro e Rede Proteomica do Rio de Janeiro, Rio de Janeiro, Brasil
gilberto@iq.ufrj.br

Wim Degrave
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Rio de Janeiro, Brasil
wdegrave@fiocruz.br

Maria da Gloria da Costa Carvalho
Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
mgccosta@biof.ufrj.br

Resumo
Este estudo apresenta nova estratégia de inferência direcionada a detectar presença de doenças em amostras biológicas. Diferencialmente dos métodos existentes, esta técnica é aplicável quando o número de patologias e as mesmas são desconhecidos. Esta é exemplificada através de software que denominamos "Máquina de Agrupamento por Elipsóide", do inglês, *Ellipsoid Clustering Machine* (ECM). O mesmo identifica regiões conservadas em perfis proteômicos obtidos por espectrometria de massa de amostras biológicas de indivíduos controles e estima limites para classificação baseando-se na variância da expressão protéica. O software também pode ser utilizado para inspeção visual de reprodutibilidade de dados. O ECM foi avaliado utilizando perfis protéicos do soro de pacientes com a doença de Hodgkin e de indivíduos controle. De acordo com a validação cruzada *leave-one-out*, o ECM separou corretamente os grupos se baseando apenas na informação de quatro picos espectrais selecionados. Este trabalho descreve o algoritmo e apresenta imagens de modelos 3D representativos da separação. O software está disponível na página do projeto na internet junto com modelos interativos e uma animação demonstrando o método.

Palavras-chave
Espectrometria de massa, aprendizagem de máquina, reconhecimento de padrões, *clustering*, doença de Hodgkin, proteômica

Introdução

Biomarcadores e a proteômica

Durante os últimos 40 anos a possibilidade de detecção precoce do câncer por biomarcadores veio para a vanguarda como promessa para transformar o diagnóstico e prognóstico médico (CONRADS et al., 2004). Entretanto, a busca por um único biomarcador que fosse específico para uma patologia não obteve sucesso até hoje. Até o antígeno prostático específico (PSA), utilizado para diagnosticar câncer de próstata quando altamente expresso em homens, poderá fornecer resultados equivocados, além de não ser específico (TROYER et al., 2004).

Um dos objetivos da proteômica é caracterizar estados de "sistemas biológicos" através de alterações no perfil de expressão de proteínas. Existem diversos métodos para obtenção de perfis proteômicos ou comparação de expressão de níveis de proteínas (ex. marcação por isótopos, comparação entre intensidade de íons, comparação entre o número de contagens espectrais (LIU et al., 2004)); a maioria das técnicas existentes fazem uso da espectrometria de massa durante a etapa de identificação de proteínas. As primeiras abordagens empregadas na "visualização" do perfil proteômico faziam uso da eletroforese bidimensional (2DE) (O´FARRELL, 1975). Esta técnica combina uma focalização isoelétrica na primeira dimensão com a eletroforese em gel de poliacrilamida contendo dodecil sulfato de sódio (SDS) na segunda. A partir daí, a técnica foi aperfeiçoada em vários aspectos como no poder resolutivo e tornando-se operacionalmente simples devido a gradientes de pH, comercialmente disponíveis, imobilizados em suportes. Dependendo das condições de fracionamento e da espécie biológica investigada, esta técnica permite observar centenas de proteínas simultaneamente. A identificação de marcadores pela técnica do 2DE consiste em contrastar diferenças no gel de indivíduos portadores da patologia em estudo em relação ao gel de indivíduos saudáveis. Após coloração por prata ou azul de coomasie, os *spots* no gel com intensidades diferentes podem representar os biomarcadores, neste caso, proteínas diferencialmente expressas isoladas de acordo com seu ponto isoelétrico e massa molecular. A limitação inicial na identificação da maioria dos sinais protéicos revelados pela segunda dimensão foi superada devido ao auxílio dos espectrômetros de massa e dos computadores. Para que uma molécula seja analisada por espectrometria de massa, ela primeiramente deve ser ionizada. Com a criação de metodologias capazes de ionizar biomoléculas sem degradá-las, a caracterização destas pelo respectivo perfil de massas, após digestão enzimática, tornou-se possível. Estes avanços fizeram com que a 2DE compusesse uma das primeiras ferramentas no arsenal das tecnologias "oma". Recentes avanços no campo da eletroforese permitiram marcar proteínas oriundas de amostras de diferentes classes com fluorófilos permitindo a separação das amostras em um único gel por técnica conhecida como DIGE (acrônimo do inglês para *differential gel electroforesis*). Certamente, entre os maiores fatores limitantes da técnica de eletroforese se ressaltam a sua dificuldade de automatização e ser extremamente laboriosa.

Em 2002, Petricoin et al., pre-fracionaram peptídeos hidrofóbicos de baixo peso molecular obtidos do soro de pacientes com câncer de ovário e traçaram perfis proteômicos utilizando a técnica SELDI-TOF MS (do inglês, *Surface Enhanced Laser Desoption Ionization – Time of Flight Mass Spectrometry*). Doravante, o termo "pico" referir-se-á ao sinal gerado por um conjunto de peptídeos de mesma m/z detectados pelo espectrômetro de massas. Baseando apenas em picos pre-selecionados, por estarem diferencialmente intensificados, Petricoin et al. afirmaram ter classificado amostras de soro de mulheres com câncer de ovário com 100% de acerto e de mulheres não afetadas pela doença com 95%. A tecnologia era inovadora e promissora, pois dispensava a separação eletroforética e permitia automatização. Mais tarde, uma abordagem semelhante foi apresentada para câncer de mama, utilizando a tecnologia SELDI com a "análise de separabilidade máxima unificada" (LI et al., 2002). Outro trabalho discriminou o câncer de próstata entre indivíduos controle utilizando árvores de decisão com técnicas *boosting* (QU et al., 2002) e métodos de estatísticas clássicos. SELDI envolve a análise de pequenos grupos de proteínas, pré-selecionados por suas propriedades de afinidade com um suporte. A depleção de proteínas poderia resultar na perda de potenciais biomarcadores e mudanças no padrão proteômico (MEHTA et al., 2003).

A compreensão dos diferentes componentes de um sistema biológico e padrões diferenciais qualitativos e quantitativos de biomolecules ainda é um desafio, mesmo com todos os avanços das tecnologias "omas". A busca por múltiplos biomarcadores é crucial na elaboração de modelos probabilísticos a serem utilizados para diagnósticos diferenciados e personalizados. A identificação de painéis de biomarcadores desafia o campo da proteômica exigindo cada vez mais sensibilidade e poder de quantificação que os das técnicas existentes (eletroforese em gel, cromatografia e espectrometria de massa). Esta problemática desafia também a ciência de inteligência artificial no setor de reconhecimento de padrões. Os dados obtidos são geralmente limitados devido ao baixo número de amostras clínicas disponíveis para pesquisa. Outros fatores limitantes são: o custo dos equipamentos e reagentes, o elevado número de parâmetros por amostra, grande variabilidade entre amostras de mesma classe, limitações na reprodutibilidade das técnicas proteômicas para detecção e quantificação simultânea de milhares de proteínas, e a falta de conhecimento de uma função de densidade de probabilidade que descreva adequadamente a variação do nível de expressão de cada proteína para o caso em estudo.

A problemática de reconhecimento de padrões

A construção de modelos matemáticos que capacitam máquinas a aprender e inferir a partir de experiências

passadas são objetos de debate e discussão filosófica. O desafio envolve a construção de máquinas capazes de aprender, com especialistas, a classificar eventos. A seguir encontra-se a formalização desta problemática.

Certo fenômeno gerará eventos x de forma aleatória e independentes de acordo com uma função densidade de probabilidade $p(x)$. Os eventos serão classificados como pertencentes a uma das k classes de acordo com especialistas. Por simplicidade, aqui $k = 2$; contudo esta formulação poderá ser generalizada para valores maiores de k, desde que cada uma das k classes, ou novas subclasses, possam ser subdividias em duas. Assume-se que o especialista realize a classificação de acordo com a função de distribuição de probabilidade condicional de classe $p(y|x)$ onde $y = \{+1, -1\}$ ($y = +1$ indica que o especialista rotulou o evento x como pertencente à classe denominada de positiva (+1), e $y = -1$ para a classe negativa (–1)). As propriedades do fenômeno que gera eventos de acordo com $p(x)$, e a regra de decisão são desconhecidas, contudo ambas as funções existem.

Seja C um conjunto de dependências funcionais $[F(x)]$ que servem como funções classificadoras para a problemática em questão. As funções podem ser representadas na forma paramétrica $F(x, \alpha)$ onde α é um parâmetro pertencente ao conjunto φ. Um valor α^* especifica a função $F(x, \alpha^*)$. O conjunto φ é arbitrário, e pode ser composto de escalares, vetores ou elementos abstratos. Todas as funções do conjunto C são funções classificadoras (i.e. elas assumem apenas o valor +1 ou –1). Observando l pares

$$x_1, y_1;\ldots;x_l, y_l$$

(sendo o evento representado por x e a classificação do instrutor y). É necessário escolher, dentre as classes de funções classificadoras $F(x, \alpha)$, a função onde a sua probabilidade de classificação difere minimamente das classificações realizadas pelo especialista.

Ou seja, o mínimo do funcional

$$E(\alpha)\sum_{k=1}^{l}\int \left(y_k - F(x,\alpha)\right)^2 p(y_k|x)p(x)dx$$

deve ser obtido para minimizar o risco esperado. Com base nos argumentos supracitados, a problemática de reconhecimento de padrões foi reduzida a minimizar o risco esperado sobre a luz de dados empíricos.

Reconhecimento de padrões e biomarcadores

Um dos maiores desafios na classificação automática de amostras biológicas está em definir métodos de aprendizagem estatística aplicável a problemas que envolvam inúmeras classes (aqui, grupo controle e cada patologia existente), e onde cada amostra possui diversas características (aqui, biomoléculas). Diversos estudos na literatura, semelhantes aos mencionados acima, são **dicotômicos**, sempre discriminando entre indivíduos controle e uma única patologia. Estas abordagens geralmente empregam algoritmo de aprendizagem supervisionada para treinar sobre conjunto de dados e selecionar um arranjo de biomoléculas com expressão diferencial para estabelecer um limite de decisão separativo entre as referidas classes. Contudo, esta estratégia de indução é limitada a duas classes e pode gerar erros caso amostras contendo patologias, cujo classificador não foi treinado a reconhecer, forem apresentada.

O objetivo destes estudos a longo prazo é o desenvolvimento de sistemas especialistas para diagnosticar automaticamente amostras biológicas. A abordagem mais bem sucedida na prática é converter um problema de classificação multiclasse em diversos problemas de classificação dicotômicos e proceder com estratégias do tipo "*one against all*" ou "*all pairs*" (MAO et al., 2005; NIIJIMA et al., 2005; XU et al., 2007). Entretanto, caso o sistema especialista estiver diante de uma classe desconhecida (uma patologia que o sistema especialista nunca foi treinado para reconhecer), o conjunto existente de classificadores binários podem não detectar corretamente a patologia e emitir erro Tipo II (classificar um paciente como saudável). Dada à existência de inúmeras patologias e que gerar erro Tipo II é o "mais indesejável" por se tratar de um diagnóstico, o desenvolvimento de heurísticas para estes casos se faz necessário.

Métodos e algoritmo

Este trabalho introduz novo racional para ser utilizado na linha de frente para classificação de perfis de amostras biológicas; diferencialmente dos métodos existentes, esta técnica é aplicável quando o número de patologias e as mesmas são desconhecidos. Nosso método é capaz de identificar conjuntos de proteínas cuja expressão permanece conservada em indivíduos controle. Nossa premissa é que algumas destas proteínas poderiam estar alteradas nos pacientes. Diferentemente da detecção de biomoléculas sobre-expressas ou suprimidas, este procedimento delineia um domínio "livre de patologia" (domínio conservado) em espaço de características e poderia servir como primeiro passo, simples e direto, para o diagnóstico de doenças.

O sistema especialista proposto primeiramente utilizaria um classificador baseado nos "domínios conservados" para avaliar a probabilidade de uma amostra desconhecida pertencer à classe de "doença". Se a amostra situa-se fora dos limites da região conservada, apenas então, o sistema especialista contaria com sua coleção de classificadores (binários) para sugerir diagnóstico. A abordagem tradicional, que aplica diversos classificadores binários onde cada um é específico a uma patologia, poderia levar a uma conclusão falsa quando diante de uma nova classe uma vez que nenhum dos classificadores foi treinado pata reconhecer a doença. Nossa abordagem poderia ter maiores chances de detectar amostras biológicas pertencentes a patologias, ainda não apresentadas ao classificador, por ser treinado a reconhecer regiões conservadas de perfis de proteínas de indivíduos controle. Portanto, nosso método poderia ser capaz de alertar quanto à presença de novas classes de patologias.

Nossos resultados consistem de uma prova de princípio dos conceitos descritos. Primeiro, nós adquirimos perfis de espectro de massa do soro de indivíduos controle e pacientes com a doença de Hodgkin. Seguidamente, nós desenvolvemos algoritmo denominado Ellipsoid Clustering Machine (ECM) tendo raízes nos conceitos previamente descritos para pesquisar as regiões conservadas nos perfis de proteína de espectros de massas dos indivíduos controle. Finalmente, o algoritmo foi avaliado pelo método de validação cruzada *leave-one-out* (LOO) para verificar caso poderia classificar corretamente entre indivíduos controle e pacientes com a doença de Hodgkin. Picos de espectros de massas que poderiam corresponder a supostos biomarcadores da doença de Hodgkin também foram rastreados utilizando o novo método de seleção de atributos acima descrito.

O conjunto de dados utilizado neste trabalho foi originado do soro de 30 pacientes com a doença de Hodgkin e de 30 indivíduos controle. O soro foi coletado no Hospital Universitário Clementino Fraga Filho do Rio de Janeiro. O diagnóstico e classificação histológica foram confirmados por um hematopatologista, de acordo com o critério da Organização Mundial de Saúde. A avaliação do paciente incluiu exames físicos, sorologia, teste para HIV, radiografia do tórax, tomografia computadorizada do tórax, abdômen e biópsia da medula óssea. O soro foi armazenado em alíquotas à -80ºC. Informações demográficas, de estágio de tumor e patológicas sobre os pacientes foram então armazenadas em um banco de dados. Todos os soros foram obtidos antes de iniciar o tratamento. Todos os indivíduos controle estavam livres do câncer baseado em seus históricos clínicos e exames físicos. Nenhuma abordagem de imagem adicional ou exame de marcador de rotina foi efetuado. Todos os participantes forneceram seu consentimento por escrito para este estudo e o projeto foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal do Rio de Janeiro.

Uma análise de gel 1D foi realizada para procurar por proteínas diferencialmente expressas evidentes e para servir como uma pré-projeção para todas as amostras (CARVALHO et al., 2005). O procedimento para aquisição de perfis de espectros de massas utilizado neste trabalho é descrito por CARVALHO et al. (2007). O conjunto de dados analisado também foi originado do trabalho citado. É importante notar que dois espectros de massas foram adquiridos para cada amostra biológica e a média de cadê pico calculada objetivando a diminuição do ruído; os dados dos espectros de massas foram então discretizados em janelas de 1Da somando valores intermediários. Doravante, vamos nos referir a estes valores discretizados como *bin(s)*. Exemplos de perfis proteômicos podem ser vistos na Figura 1.

Resultados

O algoritmo ECM foi aplicado ao conjunto de dados de perfis proteômicos para definir as "regiões conservadas" com baseado na classe de indivíduos controle, inicialmente por abordagem univariada. Seu algoritmo segue as seguintes etapas: um indivíduo é caracterizado por um vetor de *bins* onde, conforme descrito acima, cada *bin* é representativo de uma região do espectro de massa. Para cada *bin,* o seu valor correspondente é utilizado para mapear indivíduos controle em espaço de características unidimensional. Após todos os indivíduos estarem mapeados, limites de decisão são estendidos de cada ponto até que pré-determinado número de *data points* fiquem compreendido dentro dos limites iniciados por outros *data points*. Uma vez que este processo é realizado em espaço unidimensional, os limites de decisão serão compostos por "linhas". Uma lista das regiões de perfis de espectro de massas mais conservadas pode ser apontada como aquelas cujos limites de decisão tenham se estendidos o mínimo para satisfazer tal critério. Mais detalhes deste algoritmo e de seu código fonte estão disponíveis na página do projeto na internet.

Após selecionar as regiões conservadas dos perfis de espectro de massas, as fronteiras de classificação por hiper-elipsóides podem ser modeladas, agora por técnica multivariada. Isto é realizado utilizando apenas dados das regiões de espectros de massas que estavam marcados como conservadas. Deste momento em diante, o ECM mapeia os indivíduos controle em espaço de características multidimensional de acordo com as intensidades de cada *bin* selecionada. Nós notamos que agora, este espaço de características tem a mesma cardinalidade do número de *bins* selecionados e o valor de cada *bin* é utilizado para mapear o indivíduo controle de acordo com um eixo ortogonal neste espaço. Hiper-elipsóides são então originados a partir do ponto mapeado representativo de cada indivíduo controle e são estendidos, semelhantemente ao processo anterior, tendo taxa de crescimento de cada eixo proporcional à variância dos dados da respectiva dimensão. Este processo é realizado em um *loop*. O crescimento de todos os elipsóides cessa quando o centro de cada elipsóide é "envolvido", por número pré-determinado pelo usuário, de elipsóides originados de outros indivíduos controle. A classificação é realizada verificando se os dados de novos espectros localizam-se dentro ou fora dos limites do hiper-elipsóides.

De acordo com o método de validação cruzada *leave-one-out* aplicado ao nosso conjunto de dados de pacientes com a doença de Hodgkin, o ECM classificou corretamente todos os indivíduos controle e todos os pacientes com doença de Hodgkin. A Figura 2 mostra o limite de decisão criado baseado nos dados dos indivíduos controle representados por grupo de elipsóides na cor azul; os pacientes com doença de Hodgkin estão representados como pequenas esferas vermelhas. que o tamanho dos eixos dos elipsóides é proporcional à variância dos dados para cada respectiva direção. As áreas das esferas vermelhas são iguais a uma constante arbitrária utilizada para mera ilustração. Um resultado digno de nota é que em geral, as esferas vermelhas mais distantes representaram pacientes em estágio mais avançado e disseminado da doença enquanto esferas mais próximas ao grupo de elipsóides representaram pacientes em estágios iniciais da doença.

A fim de avaliar visualmente quão bem as regiões conservadas selecionadas podem discriminar indivíduos controle contra uma determinada patologia, um visuali-

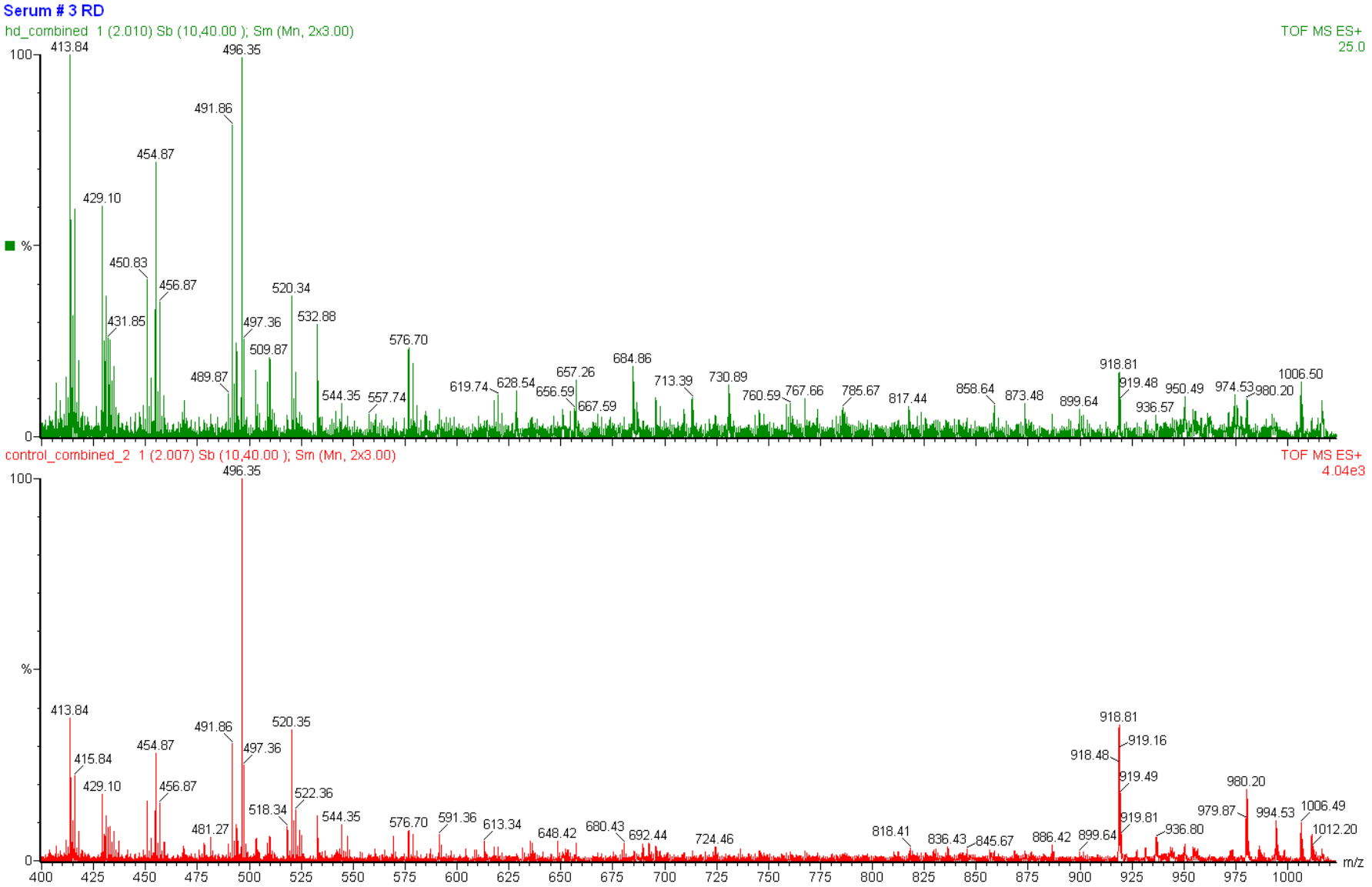

Figura 1 – Os espectros de massas verde e vermelho são exemplos de perfis proteômicos obtidos do soro de pacientes com a doença de Hodgkin e de indivíduos controle respectivamente. O eixo Y representa a intensidade de íons e o eixo X representa a razão massa / carga dos íons detectados.

zador 3D também foi disponibilizado. O visualizador é capaz de, quando trabalhando com 3 dimensões (3 *bins*), mostrar elipsóides representando sujeitos controle em azul e esferas vermelhas para representar os pacientes. O posicionamento central de cada elipsóide no espaço de características é dado pela intensidade do espectro de massas normalizada de cada respectivo biomarcador para um determinado indivíduo. O navegador de internet deve ser utilizado para visualizar o modelo VRML (Virtual Reality Modeling Language) que deve ser previamente instalado. O Cortona VRML client é sugerido uma vez que ele está disponível gratuitamente para download em [http://www.parallelgraphics.com/products/cortona/]. Modelos interativos 3D estão disponíveis na página do projeto na internet.

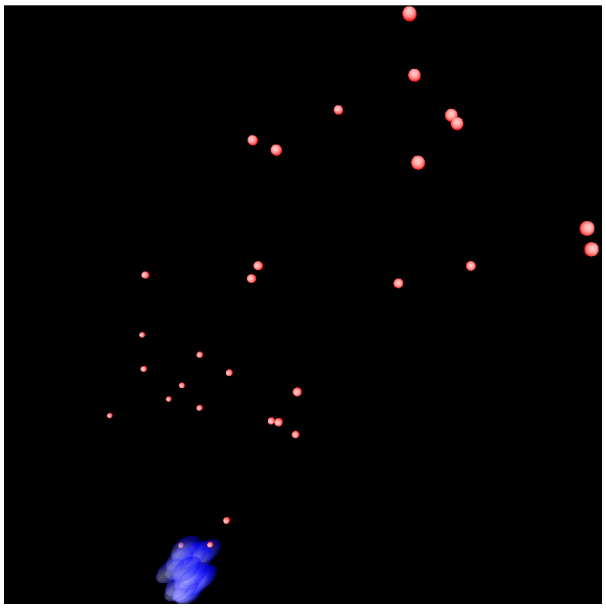

Figura 2 – Elipsóides são estendidos dos indivíduos controle mapeados em espaço de características e definem um "domínio livre da doença de Hodgkin". As esferas vermelhas representam dados dos pacientes com a doença de Hodgkin; todos eles estão localizados fora do domínio livre da doença.

Discussão e conclusões

Reconhecimento de padrões e bioinformática

Dois aspectos caracterizam os desafios de seleção de características no campo da bioinformática: a elevada cardinalidade dos dados relativos a uma amostra e limitações no tamanho do conjunto de dados a ser analisado. Para abordar estes problemas, diversos métodos de seleção de características foram desenvolvidos por especialistas nos campos de aprendizagem de máquinas e mineração de dados. Atualmente, o consenso é que não existe um método universal para a seleção de características (YANG et al., 2005); adicionalmente, sempre deve ser considerado a existência de mais de um sub-conjunto ótimo de características capazes de discriminar os dados eficientemente (YEUNG et al., 2005). Acreditamos que cada método de seleção de características possui um nicho portanto é importante o conhecimento de suas idiossincrasias, quando aplicá-lo, e estar ciente de suas limitações. Por exemplo, a resposta de um método univariado pode ser mais intuitiva de compreender por analisar cada característica independentemente. Em contrapartida, subgrupos de proteínas que possivelmente possam interagir entre si podem ser discriminados apenas por técnicas multivariadas, ao custo de maior tempo computacional.

ECM aplicado ao diagnóstico de patologias

Utilizar agrupamentos de elipsóides para definir um limite de decisão pode ser uma abordagem conservadora, mas ainda sim capaz de classificar dados que não são separáveis linearmente no espaço de características (Figure 3). O método elipsóide é capaz de encapsular e se moldar eficientemente a conjunto de dados complexos com capacidade de generalização considerável conforme sugerido na Figura 3. O método apresentado foi otimizado desde seu início para lidar com os problemas multiclasse, como nos caso de diagnósticos, mostrando um novo raciocínio sobre as abordagens de classificação tradicionais. Certamente, para testar mais profundamente e validar o método, vários tipos de patologias devem ser testados e um grande e diversificado grupo de indivíduos controle devem ser mapeados.

Ao mapear o que se supõe a ser "normal", nós imitamos um sistema imunológico artificial que similarmente, também é treinado para encontrar o que está diferente. O sistema imune aprende continuamente, uma vez que seria improvável ser concebido tendo um conhecimento a priori de todas as patologias existentes. O novo método aponta um caminho que combina espectrometria de massas com ECM para definir um domínio Euclidiano livre de câncer, baseado em limites definidos por agrupamentos de hiper-elipsóides. O ECM poderia ser interpretado como a definição geométrica de uma região sadia, e o algoritmo do ECM poderia simplesmente ser aplicado a outros campos da ciência para ser utilizado no auxilio de controle de qualidade e mapear padrões. O ECM poderia ser a base para um classificador multiclasse, oferecendo uma hipótese inicial rápida como um primeiro passo para reduzir as possíveis classes de solução. Como um segundo passo, este classificador poderia ser combinado a outros métodos (ex. SVM) para então atingir classificações binárias de maior confiança.

Disponibilidade e Requisitos

• **Nome do Projeto:** Ellipsoid Clustering Machine

• **Página do Projeto:** http://www.dbbm.fiocruz.br/labwim/bioinfoteam/templates/archives/ecm/

• **Sistema Operacional:** Plataforma independente

• **Linguagem de Programação:** Perl 5.8.6

• **Outros requisitos:** Cortona VRML é necessário para visualizar modelos interativos em 3D. Cortona pode ser obtido em http://www.parallelgraphics.com/products/cortona.

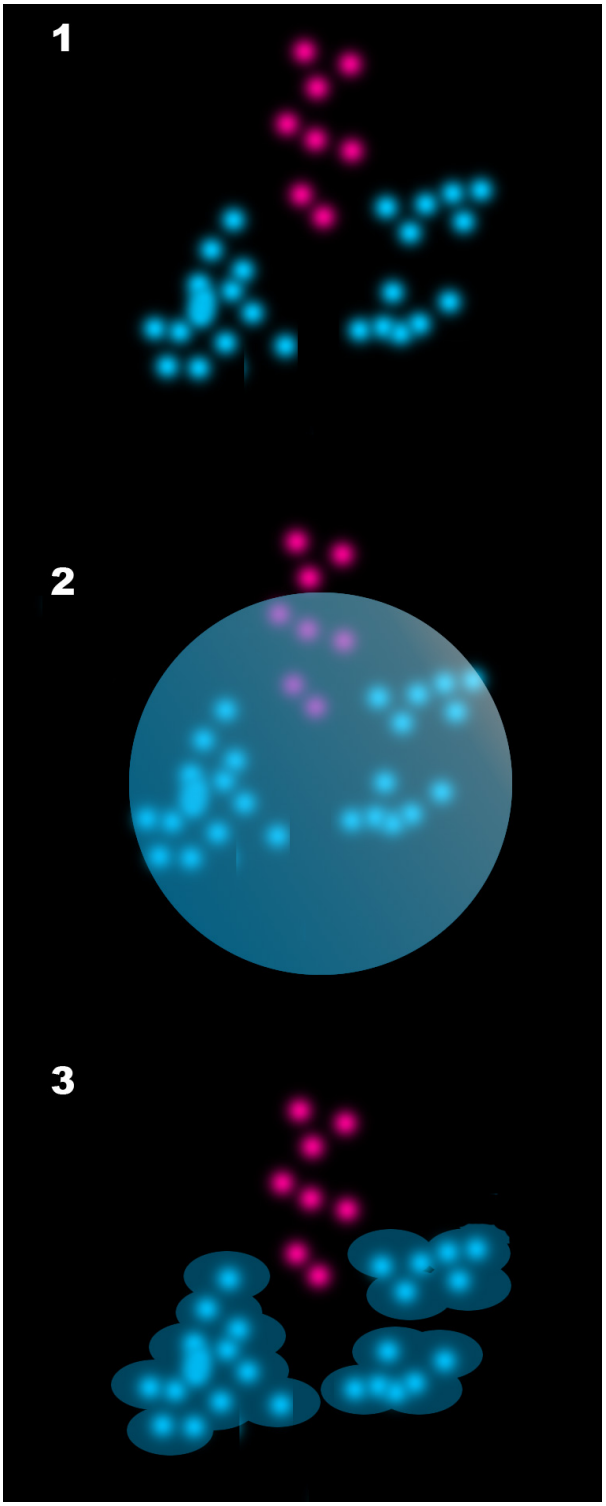

Figura 3 – Esta figura demonstra como o ECM pode ser aplicado para classificar dados que não são linearmente separáveis (1). Um único limite de decisão esférico englobando todos os indivíduos controle (pontos azuis) não separou completamente os indivíduos controle dos pacientes. Entretanto, fronteiras de decisão elípticas desenvolvidas a partir de dos dados de cada indivíduo controle podem melhor moldar-se ao conjunto de dados (3).

• **Licença:** Creative Commons Attribution-Non-Commercial-NoDerivs 2.0 License.

• O dataset do espectro de massa em formato .txt está disponível mediante solicitação aos autores.

• **Restrição à utilização por não-acadêmicos:** licença necessária.

Agradecimentos

Os autores agradecem a Fundação Ary Frauzino / Fundação Educacional Charles Darwin, Rede proteômica do Rio de Janeiro, Faperj / BBP (Cientista do Nosso Estado), CNPq, PDTIS, colaboração Fiocruz – Inca e a www.genesisdna.com.br pelo suporte financeiro.

Referências bibliográficas

CARVALHO, P.C. et al. Detection of potential serum molecular markers for Hodgkin's disease. **Jornal Brasileiro de Patologia e Medicina Laboratorial**, v.41, p.99-103, 2005.

CARVALHO, P.C. et al. Differential protein expression patterns obtained by mass spectrometry can aid in the diagnosis of Hodgkin's disease. **Journal of Experimental Therapeutics & Oncology**, v.6, p.137-145, 2007.

CONRADS, T.P. et al. Proteomic patterns as a diagnostic tool for early-stage cancer: a review of its progress to a clinically relevant tool. **Journal of Molecular Diagnostics**, v.8, p.77-85, 2004.

LI,J. et al. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. **Clinical Chemistry**, v.48, p.1296-1304, 2002.

LIU, H.; SADYGOV, R.G.; YATES, J.R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. **Analytical Chemistry**, v.76, p.4193-4201, 2004.

MAO,Y. et al. Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. **Journal of Biomedicine and Biotechnology**, v.2005, p.160-171, 2005.

MEHTA, A.I. et al. Biomarker amplification by serum carrier protein binding. **Disease Markers**, v.19, p.1-10, 2003.

NIIJIMA, S.; KUHARA, S. Multiclass molecular cancer classification by kernel subspace methods with effective kernel parameter selection. **Journal of Bioinformatics and Computational Biology**. v.3, 1071-1088, 2005.

O'FARRELL, P.H. High resolution two dimensional electrophoresis of proteins. **Journal of Biological Chemistry**, v.25, p.4007-4021, 1975.

QU, Y. et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. **Clinical Chemistry**. v.48, p.1835-1843, 2002.

TROYER, D.A. et al. Promise and challenge: Markers of prostate cancer detection, diagnosis and prognosis. **Disease Markers**, v.20, p.117-128, 2004.

XU, R.; ANAGNOSTOPOULOS, G.C.; WUNSCH, D.C. Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data. IEEE/ACM. **Transactions on Computational Biology and Bioinformatics**, v.4, p.65-77, 2007.

YANG, Y.H.; XIAO, Y.; SEGAL, M.R. Identifying differentially expressed genes from microarray experiments via statistic synthesis. **Bioinformatics**, v.21, p.1084-1093, 2005.

YEUNG, K.Y.; BUMGARNER, R.E.; RAFTERY, A.E. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. **Bioinformatics**, v.21, p.394-2402, 2005.

Sobre os autores

Paulo Costa Carvalho

Engenheiro formado pela Pontifícia Universidade Católica do Rio de Janeiro (PUC - Rio) e mestre em biologia celular e molecular com ênfase em bioinformática pelo Instituto Oswaldo Cruz onde foi orientado pelos Drs. Wim Degrave e Gilberto Barbosa Domont. Durante o mestrado, Paulo aplicou máquinas de vetor de suporte para estudar perfis proteômicos obtido do soro de indivíduos controles e pacientes com a Doença de Hodgkin. Seus interesses incluem genômica funcional, proteômica computacional, inteligência artificial, reconhecimento de padrões e computação em grade. Atualmente, Paulo é doutorando do Programa de Engenharia de Sistemas e Computação da COPPE, UFRJ sendo orientado pelo Dr. Valmir Carneiro Barbosa. Sua pesquisa abrange o desenvolvimento de métodos para extrair conhecimento de sistemas biológicos em diferentes estados quando analisados pela Tecnologia Multidimensional para Identificação de Proteínas (MudPIT).

Juliana de Saldanha da Gama Fischer Carvalho

Engenheira química pela Pontifícia Universidade Católica do Rio de Janeiro (PUC - Rio) é mestre pela Faculdade de Medicina da UFRJ onde foi orientada pela Dra. Maria da Gloria da Costa Carvalho e pelo Dr. Eduardo Marcos Paschoal. Atualmente, Juliana é doutoranda do Instituto de Química da UFRJ sendo orientada pelo Dr. Gilberto Barbosa Domont e pela Dra. Maria da Gloria da Costa Carvalho. As áreas de interesses da pesquisadora abrangem proteômica por espectrometria de massa, eletroforese diferencial em gel (DIGE) e Tecnologia Multidimensional para Identificação de Proteínas (MudPIT). Durante seu doutorado, Juliana estuda os efeitos do álcool perílico sobre glioblastoma multiforme, a forma mais agressiva dos tumores astrocíticos, em pacientes e cultura de células.