

## **UM ESTUDO DE TÉCNICAS DE CLUSTERIZAÇÃO FUZZY PARA O RECONHECIMENTO DE SPOTS EM IMAGENS DE ELETROFORESE 2D**

TAINÃ RIBEIRO CARVALHO<sup>1</sup>; MARILTON SANCHOTENE DE AGUIAR<sup>2</sup>

<sup>1</sup> Universidade Federal de Pelotas, CDTEC, Computação – [trcarvalho@inf.ufpel.edu.br](mailto:trcarvalho@inf.ufpel.edu.br)

<sup>2</sup> Universidade Federal de Pelotas, CDTEC, PPGC – [marilton@inf.ufpel.edu.br](mailto:marilton@inf.ufpel.edu.br)

### **1. INTRODUÇÃO**

Proteoma tem como definição ser a expressão de proteínas de um organismo. Nesses últimos anos, houve um aumento das pesquisas nesta área e há uma necessidade de melhoras nos métodos usados para a análise de proteoma.

Atualmente o método mais usado para análise de proteomas é a eletroforese bidimensional (2D). Este método gera uma imagem com as proteínas que compõem o proteoma, porém existem ruídos que podem ser interpretados como *spots*.

O objetivo desse trabalho é realizar um estudo sobre algoritmos de clusterização fuzzy, conceito que está contido na lógica fuzzy, como um método de separação de grupos de dados. Neste contexto, a clusterização será proposta para redução plena ou parcial dos ruídos que compõem as imagens eletroforese bidimensional.

### **2. METODOLOGIA**

O termo proteoma foi apresentado em 1995, como sendo a caracterização de um conjunto de proteínas expressas em um organismo (WILKINS et al., 1996). A definição do proteoma é semelhante à definição do genoma, o que os diferenciam são as modificações que o proteoma pode ter se houver algum estímulo interno ou externo.

Atualmente, a eletroforese bidimensional é uma das principais técnicas para análises de proteomas. O funcionamento dessa técnica se baseia na força do campo magnético, que atua nos proteomas fazendo com que as moléculas se desconectem e se espalhem sobre uma solução gelatinosa (ROCHA, 2005).

A solução é digitalizada por um *scanner* e uma imagem resultante é processada (como pode ser observada na Figura 1). Entretanto, as imagens adquiridas podem possuir ruídos, tais como poeira e até mesmo rachaduras no gel, podendo interferir no resultado final.

Em termos de aplicação, a proteômica está presente em diversas áreas de interesse como, por exemplo, na investigação de marcadores moleculares em determinadas doenças indicando a resposta da célula ou tecido a estresses externos.

Através da proteômica pode-se fazer uma comparação do perfil proteico de uma célula cancerosa com o perfil de uma célula sadia, ou de uma célula cancerosa cujo portador está sob tratamento médico, por exemplo, para entender a evolução deste tratamento (CIERO; BELLATO, 2002).

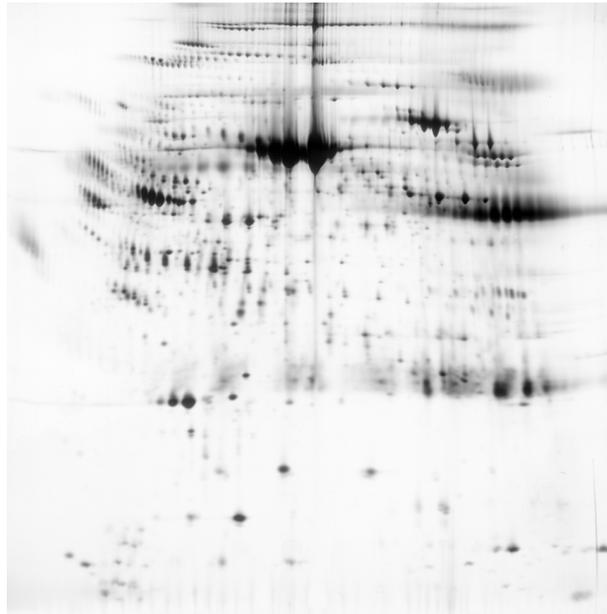


FIGURA 1 – Imagem exemplo da eletroforese 2D. Em geral, os *spots* estão representados nos pontos mais escuros e as regiões em tons de cinzas intermediários denotam os ruídos.

Uma proposta para solucionar os problemas com ruídos nas imagens de eletroforese 2D é a utilização da lógica fuzzy. Mais especificamente algoritmos de clusterização fuzzy são utilizados para a separação de dados conforme um atributo arbitrário. Dentre os métodos mais utilizados está o método Fuzzy C-Means (FCM), objeto deste estudo.

A lógica fuzzy é a utilização dos conceitos da teoria dos conjuntos fuzzy correlacionada com os conceitos da lógica tradicional, o que torna a possibilidade de resultados multivalentes, não apenas o verdadeiro e falso (KLIR, 2006). Para que haja uma relação entre uma entrada e um conjunto fuzzy é utilizada a definição de pertinência. Esta função tem como finalidade determinar um grau de pertinência para a entrada em relação ao conjunto fuzzy.

Segundo ZHERU (1996), algoritmo fuzzy c-means gera  $n$  clusters, e faz uma ligação desses clusters com os valores de amostra. Ele utiliza os valores de entrada e o grau de pertinência de cada entrada para calcular o tamanho de cada cluster.

Diferente do agrupamento *hard cluster*, que possibilita que um determinado valor  $x$  pertença apenas a um cluster, a FCM possibilita que o valor tenha um grau de pertinência em vários clusters.

Dado um conjunto de amostras  $X = \{x_1, x_2, \dots, x_n\}$ , para cada valor de entrada  $x_i$  em um cluster com centro  $c_j$ , haverá uma função de pertinência  $\mu_{ij}$ , gerando uma matriz (tabela) de funções de pertinência. Inicialmente cada função terá um valor arbitrário entre 0 e 1 e a soma deve ser igual a 1.

Após este processo, é feito um procedimento para definir cada centro de cluster. Este cálculo é dividido em 4 passos.

O cálculo de cada cluster é a divisão da soma de cada entrada  $x_i$  multiplicado pela sua respectiva função  $\mu_{ij}$  elevada a um valor de fuzzificação, relacionada ao cluster  $c_j$  pela soma de todos os  $\mu_{ij}$  elevada ao valor de fuzzificação, conforme apresentada na Equação 1.

$$c_j = \frac{\sum_{i=1}^n \mu_{ij}^m x_i}{\sum_{i=1}^n \mu_{ij}^m} \quad (1)$$

O próximo passo é calcular o valor inicial da função objeto  $J$ , que neste contexto serve como critério de parada dos cálculos dos  $c_j$ . A função objeto é a soma do produto das funções de pertinências  $\mu_{ij}$  elevada a um valor de fuzzificação  $m$  pelo quadrado da distância entre a  $i$ -ésima amostra  $x$  e o  $j$ -ésimo centro de cluster  $c$ , conforme apresenta a Equação 2 abaixo.

$$\sum_{i=1}^n \sum_{j=1}^p \mu_{ij}^m (x_i; c_j)^2 \quad (2)$$

O terceiro passo é calcular o novo valor de  $\mu_{ij}$  que é feito pela seguinte Equação 3.

$$\mu_{ij} = \frac{\left(\frac{1}{d(x_i; c_j)}\right)^{\frac{2}{m-1}}}{\sum_{k=1}^p \left(\frac{1}{d(x_i; c_k)}\right)^{\frac{2}{m-1}}} \quad (3)$$

Por fim se os critérios de parada forem satisfeitos o cálculo já está concluído, caso contrário, deve-se retornar para o passo 1. Um dos critérios de parada mais utilizados é que a distância entre a função objeto atual pela função objeto anterior seja menor que um  $\epsilon$ . Outro critério é determinar *a priori* um número máximo de iterações.

### 3. RESULTADOS E DISCUSSÃO

Neste momento, ainda não há análise que possam ser feitas porque o trabalho se encontra na etapa de revisão bibliográfica. Entretanto, vislumbra-se como próximas etapas: i) realizar um estudo amplo das técnicas visuais de análise de géis de eletroforese 2D, para que então se entenda a aptidão de cada uma e de adaptada para este trabalho; e, ii) validar os resultados alcançados, através de interações com pesquisadores do Laboratório de Bioinformática do PPGB (Programa de Pós-Graduação em Biotecnologia) da UFPel.

### 4. CONCLUSÕES

Este estudo tem como objetivo melhorar a qualidade das análises de imagens oriundas da eletroforese bidimensional. Neste sentido, busca-se estudar técnicas computacionais mais adequadas para a segmentação e classificação de imagens deste tipo.

Este trabalho está no contexto de um Projeto em desenvolvimento no grupo de pesquisa, que pretende explorar diversas possibilidades de integração entre a lógica fuzzy e algoritmos evolutivos no processo de reconhecimento dos *spots* em contraste aos métodos *crisp* (não-fuzzy) para processamento de imagens oriundas de géis de eletroforese.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS

CIERO, L.; BELLATO, C. Proteoma: avanços recentes em técnicas de eletroforese bi- dimensional e espectrometria de massa. **Biotecnologia Ciência e Desenvolvimento**, [S.l.], v.29, p.158–164, 2002.

HÖPPNER, F. Speeding up fuzzy c-means: using a hierarchical data organization to control the precision of membership calculation: **Fuzzy Sets and Systems – Clustering and modeling**, Elsevier North-Holland, Inc. Amsterdam, The Netherlands, v.128, n.3, p.365-376, 2002.

KLIR, G. J. **Uncertainty and informação: foudations of generalized information theory**. Hoboken, New Jerseu, Usa: John Wiley & Sons, 2006, p 499.

ROCHA, T.L.; SILVA, M.C.; SÁ, M.F. Eletroforese bidimensional e análise de proteomas. **Comunicado Técnico da Embrapa**,v.136,2005.

WILKINS, M. R.; PASQUALI, C.; APPEL R. D.; OU, K.; GOLAZ, O.; SANCHEZ, J. C.; JAN, J. X.; GOOLEY, A. A.; HUGUES, E.; HUMOHERY-SMITH, I.; WILLIANS, K. L.; HOCHSTRASSER, D. F. Form proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. **Nature Biotechnology**,[S.l.], v.14, p.61 – 65,1996.

ZHERU, C.; HONG, Y.; TUAN, P. **Fuzzy Algorithms : With Applications to Image Processing and Pattern Recognition**. London, Uk, Wolrd Scientific, 1996.