Aplicando Fuzzy C-Means para o Reconhecimento de Spots em Imagens Oriundas de Géis de Eletroforese Bidimensional

Geancarlo S. Maydana, Marlon da S. Dias, Marilton S. de Aguiar
Grupo de Computação Gráfica e Processamento de Imagens
Universidade Federal de Pelotas
Pelotas, Brasil
Contato: gsmaydana@inf.ufpel.edu.br, mdsdias@inf.ufpel.edu.br, marilton@inf.ufpel.edu.br

***Abstract*—Proteomics is defined as the large-scale characterization of the set of proteins expressed in a cell or tissue. Lately, two-dimensional gel electrophoresis is one of the most used techniques related to proteomics. It consists of the migration and separation of molecules, placed on a gel, according to the strength of a electric field. In order to see these proteins, it is necessary to use some kind of reagent of revelation, which ends up resulting in a two-dimensional profile of spots. Afterwards, this gel is scanned and produces an image, and then this image may be analyzed. Usually, there is noise in this kind of image. Thinking on it, this work presents a technique using fuzzy logics to find spots.**

***Keywords*-spot detection; eletrophoresis; clustering; fuzzy logic**

***Resumo*—A proteômica é definida como sendo a caracterização em larga escala do conjunto de proteínas expressas em uma célula ou tecido. Atualmente, uma das principais técnicas usadas na proteômica é a eletroforese bidimensional, baseada na separação e na migração das moléculas carregadas, sobre um gel, em função da aplicação de um campo elétrico. Estas proteínas podem ser detectadas por uma variedade de reagentes de revelação, observando-se, no final, um perfil bidimensional de pontos (spots). Esse gel é escaneado e a imagem resultante é processada. As imagens resultantes possuem grande quantidade de ruído. Neste contexto, este trabalho apresenta uma técnica fuzzy para o reconhecimento de spots.**

***Keywords*-detecção de spots; eletroforese; clusterização; lógica fuzzy**

I. Introdução

O proteoma indica as proteínas expressas em um organismo ou tecido. Enquanto o genoma representa a soma de todos os genes de um indivíduo, o proteoma não é uma característica fixa de um organismo. O proteoma altera com o estado de desenvolvimento do tecido ou mesmo sob as condições nas quais o indivíduo se encontra. Portanto, investigar diretamente os produtos dos genes é uma forma de estudar doenças e qualquer problema biológico complexo. O termo proteoma foi introduzido em 1996 para descrever as proteínas expressas em um genoma [1], [2].

A proteômica pode ser vista como uma metodologia de seleção da biologia molecular, a qual tem como objetivo documentar a distribuição geral de proteínas da célula, identificar e caracterizar proteínas individuais de interesse e principalmente elucidar as suas associações e funções. Além disso, a proteômica é uma abordagem para identificar, quantificar e estudar as modificações pós-traducionais das proteínas em uma célula, tecido ou, mesmo, organismos [3]. Por exemplo, para entender no nível molecular como uma célula funciona em um indivíduo doente e em um sadio é preciso ter conhecimento das proteínas e de outros componentes celulares que estão presentes, como eles interagem e o resultado de suas interações [3]. A análise ao nível de proteína é muito necessária, pois o estudo dos genes através do sequenciamento de DNA não pode adequadamente prever a estrutura dinâmica das proteínas, uma vez que é ao nível das proteínas que muitos dos processos de uma célula ocorrem, onde processos patológicos acontecem e aonde muitas das drogas atuam. 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 [3].

Atualmente, uma das principais técnicas usadas na proteômica é a eletroforese bidimensional, que é uma técnica de separação de moléculas baseada na migração destas moléculas numa solução, quando aplicado um campo elétrico. Em um experimento de eletroforese bidimensional em gel, as proteínas (ou outras moléculas) são separadas em duas dimensões, de modo que todas as proteínas/moléculas se espalhem por todo o gel. Na primeira etapa do experimento, as proteínas são separadas com base nas diferenças de ponto isoelétrico, sendo esta primeira etapa chamada de focalização isoelétrica (IEF). A amostra é aplicada em uma fita gelatinosa que possui um gradiente de pH e, em seguida, é submetida a um potencial elétrico. Assim, as proteínas migrarão até que o ponto isoelétrico seja alcançado, isto é, o ponto em que a carga global sobre a proteína é 0 (uma carga neutra). Neste ponto, a proteína parará de migrar, ocorrendo então a primeira

etapa de separação.

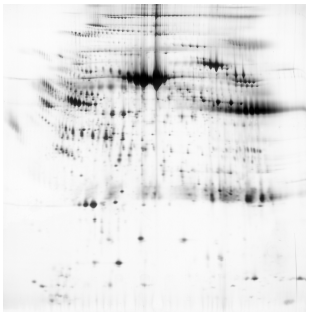
Na etapa seguinte, as proteínas são separadas pela diferença de massa molecular. A fita gelatinosa – contendo as proteínas separadas por diferença de ponto isoelétrico – é anexada a um segundo gel que permitirá a migração destas proteínas na segunda dimensão. Na segunda dimensão, um potencial elétrico é aplicado novamente, mas a um ângulo de 90 graus a partir do primeiro campo. Então, as proteínas são atraídas para o lado mais positivo do gel.

O gel atua como uma peneira molecular e, portanto, quando a corrente é aplicada, separa as proteínas com base no seu peso molecular: as proteínas maiores migram vagarosamente pelo gel ficando retidas na peneira; e,as proteínas menores se locomovem mais rapidamente depositando-se na região mais baixa do gel [4]. Como resultado deste processo, tem-se um gel com proteínas espalhadas sobre a sua superfície.

Estas proteínas podem ser detectadas por uma variedade de reagentes de revelação, sendo o nitrato de prata e o azul de Coomassie (*Coomassie Blue*) os mais utilizados. Após a coloração do gel, observa-se um perfil bidimensional de pontos (*spots*), sendo que em cada ponto há múltiplas cópias de uma proteína. Por fim, este gel é escaneado e a imagem resultante pode ser processada [4].

As imagens de eletroforese bidimensional podem conter ruídos, bem como partículas de poeira e, até mesmo, rachaduras no gel, e isso pode interferir no resultado final da análise de reconhecimento dos *spots* [5]. A Figura 1 mostra exemplos de imagens de géis oriundas de eletroforese bidimensional que ilustram esse problema. Pode ser observado na Figura 1b a presença de uma quantidade considerável de ruído, na forma de borrões, o que torna mais complexa a tarefa de determinar os *spots*.

(a) Imagem com quantidade moderada de ruído



(b) Imagem com grande quantidade de ruído

Figura 1. Imagens de géis de eletroforese bidimensional

II. Reconhecimento Fuzzy de Padrões

Os conjuntos fuzzy são o modelo mais tradicional para o tratamento de informações vagas e inexatas. Introduzido por Zadeh em [6] tem como objetivo permitir um elemento pertencer, com mais ou menos intensidade, a uma dada classe. A representação com conjuntos fuzzy utiliza conjuntos para representar a informação que não é precisa e emprega lógica fuzzy para a tomada de decisão [7], [8], provendo um mecanismo para representar e manipular algum tipo de incerteza e ambiguidade. Pode-se observar que algoritmos fuzzy têm várias aplicações principalmente em análise de imagens e reconhecimento de padrões [9].

Em geral existem duas classes de técnicas de reconhecimento de padrões: os métodos supervisionados e os não-supervisionados. Em um método supervisionado, tem-se um conjunto de amostras de treinamento em diferentes classes. Para estes dados de treinamento, precisa-se encontrar uma função de mapeamento, ou construir um classificador, que pode ser um conjunto de regras fuzzy, uma rede neural, uma árvore de decisão, ou simplesmente um conjunto de Equações matemáticas, tal que as amostras sejam classificadas corretamente em suas classes.

Em métodos não supervisionados, não tem-se amostras de treinamento, e em alguns casos, não se sabe o número exato de classes. Neste caso tem-se um conjunto de dados não-rotulados:

$$x_1, x_2, \ldots, x_n \tag{1}$$

A tarefa é dividir estes dados em vários grupos de acordo com uma medida de similaridade ou estrutura inerente dos dados. Isto pode ser feito por usar um procedimento de agrupamento. Em um procedimento de agrupamento *hard*, uma amostra é atribuída a um grupo ou não. Já em um procedimento de agrupamento fuzzy, uma amostra é atribuída a cada um dos grupos de acordo com uma função de pertinência. Os valores de pertinência desempenham um importante papel no processo de agrupamento e tornam o processo de classificação mais flexíveis e robustos ao lidar com dados incertos e ruidosos.

A. Algoritmo Fuzzy C-means

No processo *crisp*, cada amostra é atribuída a somente um cluster e todos os clusters são conjuntos disjuntos. Na prática, entretanto, existem muitos casos em que os clusters não são completamente disjuntos e os dados podem ser classificados como pertencendo mais a um cluster que a outro.

Tal situação não pode ser atendida por um processo de seleção *crisp*. Portanto, a separação dos *clusters* traz uma noção *fuzzy*, e as representações de estruturas de dados reais podem então serem tratadas com mais precisão por métodos de agrupamento *fuzzy*. Nestes casos, é necessário descrever a estrutura de dados em termos dos *clusters fuzzy*.

O algoritmo Fuzzy C-means (FCM) é o mais conhecido e a técnica de agrupamento *fuzzy* mais utilizada. Este algoritmo é desenvolvido baseado na minimização iterativa da seguinte função critério [10], [11], [12]:

$$J(U,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \mid x_k - v_i \mid^2 \tag{2}$$

onde

- $x_1, \ldots, x_n$ são n vetores de dados de amostra;
- $V = \{v_1, \ldots, v_c\}$ são os centros dos *clusters*;

- $U = [u_{ik}]$ é uma matriz $c \times n$, onde u_{ik} é o i-ésimo valor de pertinência da k-ésima amostra de entrada x_k, e os valores de pertinência satisfazem as seguintes condições

$$0 \leq u_{ik} \leq 1 \qquad i = 1, 2, \ldots, c;\ k = 1, 2, \ldots, n \quad (3)$$

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad k = 1, 2, \ldots, n \quad (4)$$

$$0 < \sum_{k=1}^{n} u_{ik} < n \qquad i = 1, 2, \ldots, n \quad (5)$$

- $m \in [1, \infty)$ é um fator de peso expoente.

A função objetivo é a soma das distâncias Euclidianas entre cada amostra e seu centro de cluster correspondente elevada ao quadrado, com as distâncias sendo ponderadas pela pertinência fuzzy.

O algoritmo é iterativo e faz uso das seguintes Equações [13]:

$$v_i = \frac{1}{\sum_{k=1}^{n} u_{ik}^m} \sum_{k=1}^{n} u_{ik}^m x_{ik} \qquad i = 1, 2, \ldots, c \quad (6)$$

$$u_{ik} = \frac{\left[\frac{1}{|\, x_k - v_i \,|^2}\right]^{1/(m-1)}}{\sum_{j=1}^{c} \left[\frac{1}{|\, x_k - v_j \,|^2}\right]^{1/(m-1)}} \qquad i = 1, 2, \ldots, c;\ k = 1, 2, \ldots, n \quad (7)$$

Para calcular um centro de cluster, todas as amostras de entrada são consideradas e as contribuições das amostras são ponderadas pelos valores de pertinência. Para cada amostra, seu valor de pertinência em cada classe depende de sua distância ao centro de cluster correspondente. O fator m reduz a influência de pequenos valores de pertinência. Quanto maior o valor de m, menor é a influência de amostras com menor valor de pertinência.

O FCM consiste dos seguintes passos [13]:

1) Inicialize $U^{(0)}$ aleatoriamente ou baseado em uma aproximação; inicialize $V^{(0)}$ e calcule $U^{(0)}$. Defina o contador de iteração $\alpha = 1$. Selecione o número de centro de classes c e escolha o peso m.
2) Calcule os centros dos *clusters*. Tendo $U^{(\alpha)}$, calcule $V^{(\alpha)}$ de acordo com a Equação 6.
3) Atualize os valores de pertinência. Tendo $V^{(\alpha)}$, calcule $U^{(\alpha)}$ de acordo com a Equação 7.
4) Pare se
$$\max |\, u_{ik}^{(\alpha)} - u_{ik}^{(\alpha-1)} \,| \leq \varepsilon \quad (8)$$
senão faça $\alpha = \alpha + 1$ e vá para o passo 2, onde ε é um número pequeno pré-especificado representando a menor mudança aceitável em U.

III. Modelo Proposto

O modelo proposto visa fazer a clusterização das imagens com o intuito de encontrar *spots*, utilizando o algoritmo de clusterização Fuzzy C-means. A análise é feita com base na imagem provida ao final do exame, onde as moléculas já sofreram os efeitos dos campos elétricos e pararam de se mover, como exemplo pode-se citar as imagens mostradas anteriormente na Figura 1. Partindo da imagem, os únicos dados presente são os tons de cinza que compõem a imagem. Baseando-se nos princípios da clusterização, acredita-se que os *spots* apresentam tons de cinza similares e, assim, pertençam ao mesmo cluster.

Testes preliminares foram executados em porções menores de imagem, pois as imagens completas de eletroforese bidimensional podem chegar a aproximadamente 1000×1000 pixels. A Figura 2 demonstra a clusterização em uma imagem reduzida, 54×54 pixels. A Figura 2a é a imagem original. Ela representa uma parte da imagem maior. As cores na Figura 2a foram invertidas, é o negativo da imagem apresentada na Figura 1. Nas Figuras 2b e 2c mostra o resultado da Figura 2a após a clusterização. Cada cor nas imagens representam um cluster.

(a) Porção da imagem de eletroforese bidimensional

(b) Imagem clusterizada, utilizando 3 clusters

(c) Imagem clusterizada, utilizando 4 clusters

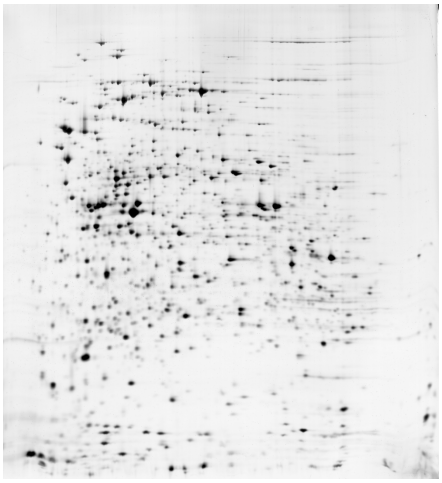
Figura 2. Classificação em uma porção da imagem de eletroforese bidimensional.

Em [14] é apresentado uma base de imagens oriundas de eletroforese bidimensional. As imagens presentes nessa base de dados já foram testadas e apresentam *spots* confirmados, os quais foram usados para medir a precisão do modelo. Logo, há uma lista informando os pontos na imagem que representam um *spot*. Os pontos apresentados nessa lista são *spots* confirmados, entretanto, ainda há outros pontos na imagem que não foram confirmados *spots*.

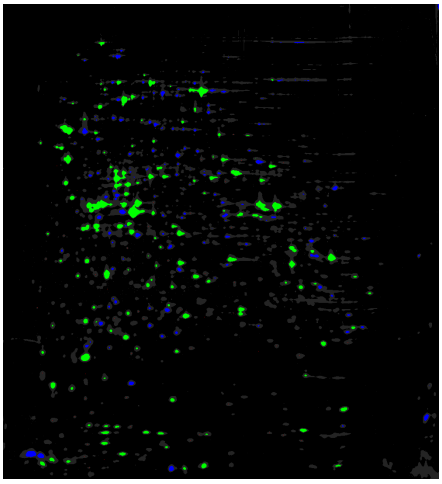
Diversos testes foram realizados com diferentes configurações do modelo. O resultado mais promissor foi utilizando três clusters. Com base nessa configuração, observa-se os clusters representando o fundo da imagem, redondezas ou rastros de um *spot* e, por fim, os *spots*. A Figura 3 mostra um teste realizado com uma das imagens da base utilizando essa configuração. A Figura 3a representa a imagem original, chamada de ECOLI. Os três clusters, mostrados na Figura 3b, são representados pelas cores preto (cluster 1); cinza (cluster 2); e azul e verde (cluster 3). O cluster 3 representa as áreas na imagem consideradas *spots*.

O modelo informa o grau de pertinência de cada ponto a um dos clusters. A função máximo é aplicada como método de defuzzificação, ou seja, o cluster do ponto é designado com base no maior grau de pertinência entre clusters. Assim, tem-se a informação do cluster de cada ponto. Para poder contar um *spot* é necessário saber a área que representa uma *spot*, o conjunto de pontos próximo pertencentes ao cluster que representa os *spots*.

Com o objetivo de formar e contar os *spots*, um algoritmo de preenchimento é usado. Parte-se de um ponto que é *spot*



(a) Imagem original - ECOLI



(b) Imagem após clusterização

Figura 3. Imagens de eletroforese bidimensional utilizadas para clusterização.

e dele verifica-se a vizinhança. Os vizinhos que também são *spots* serão os próximos a serem verificados. Esse processo se repete até não haver vizinhos classificados como *spot*. Com isso, todos esses pontos avaliados são considerados um *spot*. Esse processo é aplicado a toda imagem.

O cluster 3 possui duas cores, ambas representam *spots* encontrados pelo modelo. A diferença é que as áreas verdes são *spots* confirmados por um especialista em [14]. Os azuis não foram confirmados, mas isso não diz diretamente que eles não possam ser *spots*. Na verdade, são áreas que poderiam ser consideradas *spots*, indicando que elas precisam de uma análise mais detalhada. De acordo com o protocolo deste exame, pode ser necessário recortar uma porção do gel para uma análise mais detalhada e, assim, esta região em que o modelo indica ter spots não comprovados poderia ser uma sugestão de recorte. Os pontos vermelhos, são *spots* confirmados em [14] que não foram encontrados pelo modelo. No total, a lista de *spots* confirmados contém 206 *spots*. O modelo encontrou 279 *spots*, onde 151 estão na lista de confirmados, e 55 estão faltando, resultando em uma precisão de $73,3\%$.

IV. Considerações Finais

Imagens de eletroforese bidimensional são difíceis de analisar. Há o problema de possíveis ruídos na imagem, o que pode dificultar na análise. Porém, além disso, os *spots* podem apresentar características bem diferentes. Por exemplo, dos *spots* confirmados na lista, mas não encontrados pelo modelo, apresentam tons de cinza muito distintos. A maioria dos confirmados possui um tom mais forte, entretanto, alguns dos confirmados possuem tons de cinza bem baixos. Isso é um grande desafio a ser superado pela clusterização.

Como trabalhos futuros, é previsto a utilização de algumas técnicas de pré-processamento. É possível utilizar algumas técnicas como remoção do fundo, normalização, equalização e segmentação da imagem. As imagens utilizadas são relativamente grandes, contendo muita informação. A primeira técnica tem o intuito de reduzir os dados a serem processados, o que pode acelerar o processo. A segunda e terceira, visam mudar os dados apresentados, podendo influenciar diretamente na maneira com que a função fuzzy os clusteriza. E, por fim, a última técnica, a segmentação da imagem, tem como objetivo dividir a imagem, resultando na clusterização de imagens menores. A imagem original apresenta regiões com intensidades diferentes, com isso pretende-se achar aqueles *spots* com tons de cinza mais discrepantes.

Referências

[1] M. R. Wilkins, C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J. C. Sanchez, J. X. Jan, A. A. Gooley, E. Hughes, I. Humohery-Smith, K. L. Willians, and D. F. Hochstrasser, "From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis," *Nature Biotechnology*, vol. 14, pp. 61–65, 1996.

[2] N. G. Anderson and N. L. Anderson, "Twenty years of two-dimensional electrophoresis: past, present and future," *Electrophoresis*, vol. 17, pp. 443–453, 1996.

[3] L. Ciero and C. Bellato, "Proteoma: avanços recentes em técnicas de eletroforese bidimensional e espectrometria de massa," *Biotecnologia Ciência e Desenvolvimento*, vol. 29, pp. 158–164, 2002.

[4] T. L. Rocha, M. C. Silva, and M. F. Sá, "Eletroforese bidimensional e análise de proteomas," *Comunicado Técnico da Embrapa*, vol. 136, 2005.

[5] M. A. Savelonas, E. A. Mylona, and D. Maroulis, "Unsupervised 2d gel electrophoresis image segmentation based on active contours," *Pattern Recogn.*, vol. 45, no. 2, pp. 720–731, 2012.

[6] L. A. Zadeh, "Fuzzy sets," *Information Control*, vol. 8, pp. 338–353, 1965.

[7] ——, "A theory of approximate reasoning," *Machine Intelligence*, vol. 9, 1979.

[8] ——, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy Sets and Systems*, vol. 100, pp. 9–34, 1999.

[9] I. Bloch and H. Maître, "Fuzzy mathematical morphology," *Ann. Math. Artif. Intell*, vol. 10, pp. 55–84, 1993.

[10] J. C. Bezdek and S. K. Pal, Eds., *Fuzzy Models for Pattern Recognition*, ser. A selected reprint volume. IEEE Press, 1992, methods That Search for Structures in Data.

[11] H. Zimmermann, *Fuzzy set theory and its applications*. Kluwer Academic, Dordrecht, 1985.

[12] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

[13] Z. Chi, H. Yan, and T. Pham, *Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition*, ser. Advances in fuzzy systems - applications and theory. World Scientific, 1996.

[14] C. Hoogland, K. Mostaguir, J.-C. Sanchez, D. F. Hochstrasser, and R. D. Appel, "Swiss-2dpage, ten years later," *Proteomics*, vol. 4, no. 8, pp. 2352–2356, 2004.