Reconhecimento de Padrões em Dados de Expressão Gênica de Pacientes Portadores de Osteogênese Imperfeita

Diogo Pereira Silva de Novais[1], **Carla Martins Kaneto**[1], **Paulo Eduardo Ambrósio**[1]

[1]Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia
Universidade Estadual de Santa Cruz (PPGMC - UESC)
Ilhéus - Bahia - Brasil

Abstract. *The growing production of molecular data made possible by advances in laboratory technologies has motivated research related to analysis supported by mathematical and computational models. This paper discusses the use of clustering algorithms for gene expression data analysis of patients with Osteogenesis Imperfecta, which aims to assess the relationship between the grounds of the models and biological relevance of clusters obtained by different algorithms.*

Resumo. *A crescente produção de dados biomoleculares possibilitada pelos avanços das tecnologias laboratoriais tem motivado pesquisas relacionadas à análise apoiada por modelos matemáticos e computacionais. Este trabalho discute a utilização de algoritmos de agrupamento para análise de dados de expressão gênica de pacientes com Osteogênese Imperfeita, no qual pretende-se avaliar a relação entre a fundamentação dos modelos e relevância biológica dos agrupamentos obtidos por diferentes algoritmos.*

1. Introdução

A análise e simulação computacional de dados biológicos teve como origem de suas demandas a evolução das técnicas de análises biomoleculares laboratoriais, que resultaram na produção de grandes massas de dados que atualmente são compartilhadas em bancos de dados especializados, mantidos e compartilhados pela comunidade científica. Um grande exemplo de projeto possibilitado por essas novas técnicas laboratoriais é o sequenciamento completo do genoma humano [Dougherty 2005].

Uma das áreas que tem ganhado representatividade na área de Reconhecimento de Padrões em Bioinformatica é a análise de dados de expressão gênica em pacientes portadores de doenças específicas, visando a descoberta de relações entre os genes estudados ou para descoberta de marcadores biomoleculares para estas doenças.

A Osteogênese Imperfeita é uma doença caracterizada por uma desordem genética, que entre outros fatores, leva a uma produção defeituosa ou insuficiente de colágeno, tendo como consequência uma osteopenia generalizada que gera problemas como baixa estatura, fragilidade óssea excessiva e em quadros mais graves, a morte de portadores da doença [Kaneto 2011].

Com a evolução das tecnologias de análise biomolecular surgiram os microarranjos ou chips de DNA, que possibilitam a análise de perfis de expressão automática de centenas, ou até milhares de genes simultaneamente. A análise é feita através de circuitos controlados computacionalmente, o que permite a análise paralela de várias amostras, com uma boa precisão. O resultado de uma análise de microarranjo

é uma tabela contendo o nível de expressão de cada gene para cada uma das amostras [Lodish et al. 2003].

Um trabalho realizado por [Kaneto 2011] traz uma análise de expressão gênica durante várias etapas da diferenciação osteogênica de celulas mensequimais estromais de medula óssea de pacientes portadores de Osteogênese Imperfeita, apresentando padrões de expressão diferencial entre alguns genes relacionados a este processo, para estes pacientes em relação a amostras de indivíduos sadios. Neste trabalho são analisadas amostras de medula de doadoes normais, de um paciente com Osteogênese Imperfeita Tipo I e outro com Osteogênese Imperfeita Tipo III.

Dessa forma, este trabalho propõe avaliar o comportamento dos dados de expressão gênica ora analisados em [Kaneto 2011] com outros algoritmos de agrupamento de modo a viabilizar uma análise comparativa da relevância biológica dos grupos obtidos, possivelmente revelando relações não observáveis nos agrupamentos já realizados. Além disso, prentende-se realizar agrupamentos de dados de expressão de microRNAs, originários das mesmas amostras das quais foram obtidos os dados de expressão gênica.

A aplicação dos algoritmos no mesmo conjunto de dados, pode contribuir com o estudo de aplicação dos algoritmos estudados em dados de expressão gênica, bem como com o conhecimento molecular da Osteogênese Imperfeita, fornecendo novas relações não perceptíveis nas análises já realizadas, podendo gerar novas demandas de pesquisa relacionadas aos genes e microRNAs analisados.

2. Algoritmos de Agrupamento

Os algoritmos de agrupamento têm por objetivo a descoberta de padrões em conjuntos de dados, permitindo seu particionamento em grupos, de modo a revelar similaridades e diferenças entre padrões, possibilitando a inferência de conclusões sobre o objeto pesquisado [Theodoridis and Koutroumbas 2003].

Entre as mais comuns aplicações de algoritmos de agrupamento estão a redução de dimensionalidade, que permite a análise de grupos ao invés da análise de dados de altas dimensões, a geração de hipóteses, a confirmação de hipótese de fenômenos estarem relacionados, e a predição de comportamento a partir do grupo ao qual pertence determinado padrão [Theodoridis and Koutroumbas 2003].

Dentre os diversos algoritmos de agrupamento conhecidos, são apresentados a seguir três importantes algoritmos que são amplamente utilizados no reconhecimento de padrões e agrupamento de dados de biologia molecular. A escolha destes modelos foi realizada com base nas aplicações encontradas na literatura atual para análise de dados de expressão gênica, objetivo deste trabalho. Mais detalhes acerca da análise computacional de dados de expressão gênica podem ser encontrados em [Quackenbush 2001].

2.1. K-means

O k-means é um algoritmo de agrupamento muito utilizado em diversas áreas, inclusive em análises genéticas e biomoleculares, no qual deve-se conhecer a *priori* a quantidade de grupos existentes nos dados.

O algoritmo funciona de maneira iterativa, onde centroides, criados inicialmente

de maneira aleatória são reajustados em direção ao centro dos grupos que são reorganizados a cada iteração.

Cada padrão analisado define em cada iteração como seu exemplar o centróide que minimiza a função de distância escolhida. No fim da iteração, os centróides são redefinidos como a média dos vetores do grupo do qual é exemplar. O algoritmo se repete até que os centróides não sejam mais reajustados ou assim que atenda ao critério de parada estabelecido.

O k-means é constantemente utilizado em pesquisas na área de Genética e Biologia Molecular, dada sua simplicidade computacional e disponibilidade em grande parte das ferramentas estatísticas e de aprendizado computacional disponíveis. Neste trabalho, a análise dos dados e estudo do K-means se justificam por fornecer um referencial para comparação dos resultados obtidos através da aplicação de outros algoritmos nos mesmos dados utilizados em [Kaneto 2011].

2.2. Mapas Auto-Organizáveis

Os mapas auto-organizáveis, também conhecidos por redes SOM (do inglês - *Self Organizing Maps*) são um tipo de Rede Neural Artificial, com aprendizado não supervisionado utilizado no agrupamento de dados, propostos inicialmente por Kohonen em 1982 [Haykin 1999].

O principal objetivo de um SOM é a transformação de um conjunto de dados pertencentes a um espaço de dimensão arbitrária em um mapa discreto de baixa dimensionalidade, geralmente uma ou duas dimensões, de maneira topologicamente ordenada [Haykin 1999]. Com base nestas características, os SOMs geralmente possuem dois tipos de aplicações: Compressão de dados (ou redução de dimensionalidade) e a disposição de dados de modo a evidenciar semelhanças entre dados agrupados.

Em relação ao seu funcionamento, os SOMs fazem parte das redes de aprendizado competitivo, ou seja, ao ser fornecida uma entrada, todos os neurônios avaliam esta entrada com uma função discriminante e aquele que obtiver a maior avaliação para a entrada tem seus pesos ajustados. Diferente de algoritmos de aprendizado competitivos, conhecidos como *winner takes all* (em uma tradução livre, ”vencedor leva tudo”), em que apenas o neurônio com maior avaliação para entrada tem seus pesos ajustados, nos SOMs, os neurônios em uma vizinhança do neurônio ”vencedor” tem seus pesos ajustados proporcionalmente à sua proximidade do neurônio vencedor, fazendo com que grupos vizinhos no mapa discreto possuam padrões semelhantes no espaço de entrada [Haykin 1999].

Além de aplicações na área de reconhecimento de padrões em outras áreas como Reconhecimento de Fala e Processamento de Imagens, os SOMs também são utilizados para reconhecimento de padrão em dados biológicos.

2.3. Affinity Propagation

Alternativamente aos algoritmos citados anteriormente, o trabalho de Frey [Frey and Dueck 2007] propõe um algoritmo de agrupamento não hierárquico, no qual não é necessária a escolha a *priori* da quantidade de grupos. Este algoritmo é chamado de *Affinity Propagation*.

No *Affinity Propagation*, cada ponto (padrão) é visto como um nó em uma rede e todos pontos inicialmente são vistos como possíveis exemplares de um grupo. Os nós da

rede trocam mensagens buscando escolher o melhor candidato a representante do grupo ao qual pertence. As trocas de mensagens entre dois pontos, baseadas em funções de similaridade, indicam a afinidade de um ponto em escolher outro como exemplar. Através das iterações do algoritmo, exemplares são definidos, formando grupos com os elementos que o elegeram como representante [Frey and Dueck 2007].

Além de serem apresentadas nas propostas iniciais do trabalho em [Frey and Dueck 2007] aplicações na área de genética, outros trabalhos utilizam o *Affinity Propagation* para análise de dados biológicos.

3. Andamento do Trabalho

Uma vez que a proposta do trabalho envolve a utilização algoritmos de agrupamento adicionais com o objetivo de obter agrupamentos biologicamente mais significativos, os dados extraídos e pré-processados das pesquisas que o precedem, serão utilizados, de modo a possibilitar análises comparativas dos resultados obtidos por estes algoritmos.

Uma vez revisada a literatura acerca do tema e organizados os dados de expressão gênica, as seguintes etapas devem ser realizadas para conclusão do trabalho:

- Construção de Agrupamentos com os Mapas Auto-Organizáveis e com o *Affinity Propagation*;
- Análise comparativa da significância dos agrupamentos com base em evidências biológicas de semelhança entre os genes agrupados;
- Definição de Estratégia de Análise para dados de miRNA com base nos resultados obtidos para mRNA;
- Análise comparativa da significância dos agrupamentos com base em evidências biológicas de semelhança entre os miRNA agrupados ou seus genes alvo, caso conhecidos.

Referências

Dougherty, E. R. (2005). The fundamental role of pattern recognition for gene-expression/microarray data in bioinformatics. *Pattern Recognition*, 38(12):2226–2228.

Frey, B. J. and Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, 315(February):972–976.

Haykin, S. (1999). *Neural Networks: A comprehensive foundation*. Prentice-Hall, New Jersey, 2 edition.

Kaneto, C. M. (2011). *Análise da Expressão Gênica durante a diferenciação osteogênica de células mesenquimais estromais de medula óssea de pacientes portadores de Osteogênese Imperfeita*. Tese de doutorado, Faculdade de Medicina de Ribeirão Preto/USP.

Lodish, H., Arnold Berk, Matsudaira, P., Keiser, C. A., Krieger, M., Scott, M. P., Zipursky, L., and James Darnell (2003). *Molecular Cell Biology*. W. H. Freeman, 5 edition.

Quackenbush, J. (2001). COMPUTATIONAL ANALYSIS OF MICROARRAY DATA. *Nature reviews. Genetics*, 2(June):418–427.

Theodoridis, S. and Koutroumbas, K. (2003). *Pattern Recognition*. Elsevier, San Diego, 2 edition.