III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

DETEÇÃO DE GRUPOS DE OBSERVAÇÕES ATÍPICAS: UMA APLICAÇÃO EM DADOS GENÓMICOS

Ana Tavares[1], Vera Afreixo[1] e Paula Brito[2]

[1]Universidade de Aveiro
[2]Universidade do Porto

RESUMO

Este trabalho aborda o problema da deteção de grupos de observações que se afastam da maioria. O objetivo é a deteção de grupos de palavras genómicas cujo padrão de distribuição, ao longo do genoma, se distinga da maioria dos padrões. O método proposto aplica técnicas de classificação hierárquica para identificar classes de pequena dimensão, pois é nesses grupos que se espera encontrar as observações que se demarcam das restantes. Assim, a dimensão das classes serve de ponto de partida para a identificação de potenciais observações atípicas. Num segundo passo, as observações são comparadas com as restantes observações do seu grupo, por forma a avaliar a similaridade entre as distribuições. Para esse efeito é utilizada uma medida de atipicidade funcional que privilegia a forma das distribuições e não apenas a magnitude dos seus valores.

Palavras chave: Deteção de *outliers*, Classificação hierárquica, Distribuições de distâncias, Palavras genómicas.

1. INTRODUÇÃO

O ADN pode ser representado por uma sequência linear composta por quatro símbolos distintos (A, C, G, T). Um segmento de k símbolos consecutivos é designado por palavra genómica. Algumas palavras têm uma função biológica bem definida e muitas regiões funcionalmente importantes do genoma podem ser reconhecidas através da identificação de padrões de sequência, ou "motivos"[5]. Por exemplo, o trinucleótido ATG serve como um local de iniciação nas regiões de codificação (um marcador onde a tradução para a proteína começa) [6]. Conjetura-se que os padrões que têm algum significado funcional ou estrutural estão sujeitos a pressões de seleção positivas (ou negativas) durante a evolução e, consequentemente, têm uma frequência maior (ou menor) do que a esperada [2]. Isso torna a identificação de palavras genómicas e dos seus padrões de distribuição um objeto relevante de investigação.
A análise de sequências de ADN é um domínio de pesquisa amplo e alvo de várias e novas abordagens. Uma dessas abordagens é o estudo das distribuições de distância entre as palavras genómicas [1, 11]. A distância entre palavras (iguais) é definida como a diferença entre a posição dos primeiros símbolos de ocorrências consecutivas da palavra. Por exemplo, as

distâncias entre as palavras $w = AC$ no segmento $ACTGACAGGACAC$ são (4,5,2). A distribuição de distâncias de w traduz a frequência de ocorrência de cada distância, isto é, o número de vezes que a palavra se repete, a uma distância específica, na sequência de ADN em estudo.
Há palavras que revelam, ao longo do genoma, padrões de distribuição que se distinguem da maioria dos restantes, parecendo que foram geradas por um mecanismo diferente. Um exemplo bem documentado na literatura é do dinucleótido CG que, apesar de sub-representado no genoma humano, se aglomera densamente em determinadas regiões (as ilhas de CpG), revelando um perfil de distribuição bem distinto do dos restantes nucleótidos. Por outro lado, existem evidências de que várias palavras genómicas exibem um comportamento específico, que pode ser interpretado como uma assinatura da própria palavra. Palavras com sub-estruturas comuns poderão apresentar padrões de distribuição semelhantes (e.g. ACGCG e CGCGT). Existem motivos que apresentam alguma especificidade funcional, estrutural ou regulatória que são compostos por dezenas de nucleótidos podendo dar origem a um conjunto de palavras de tamanho substancialmente menor que tenham comportamentos semelhantes e bastante específicos. Deste modo, espera-se que, a existir padrões atípicos, estes sejam descritos por um pequeno grupo de palavras e não apenas por uma palavra isolada. Sugere-se que estas palavras são candidatas a apresentar algum tipo de enriquecimento funcional.
Neste trabalho, propomos uma abordagem que recorre a métodos de classificação hierárquica para detetar grupos de pequena dimensão onde se espera encontrar distribuições que se distingam da maioria e sejam, simultaneamente, muito semelhantes entre si.

2. MÉTODO

O método é composto por dois passos, um que agrupa as distribuições de acordo com a sua tendência global; e outro que aplica aos elementos de cada grupo um método de deteção de *outliers* de dados funcionais.
As distribuições empíricas podem apresentar variações que se devem a diferentes causas: variabilidade amostral, mudanças na tendência ou picos de frequência. De modo a evidenciar a tendência global da distribuição é desejável a redução da pequena variabilidade, mantendo a variabilidade mais forte. As distribuições são suavizadas com vista a uma redução da variabilidade amostral, tornando mais evidente a tendência global de cada distribuição. A suavização pode ser considerada uma aproximação à regressão não paramétrica, pelo que não exige pressupostos para a sua utilização. Neste trabalho aplicamos a suavização por funções *spline* cúbicas [9], método bastante flexível na identificação da relação funcional entre as variáveis. Para superar a arbitrariedade da escolha do valor do parâmetro que governa o equilíbrio entre a suavidade da curva e sua proximidade com os valores observados utiliza-se *cross-validation* [4].
A similaridade entre os dados suavizados é explorada através de um método de classificação hierárquica (aglomerativo). As classes de menor dimensão que são agregadas em níveis superiores do dendrograma estão potencialmente associadas aos grupos de interesse. A distância entre as classes formadas em cada etapa do procedimento aglomerativo é avaliada segundo o critério da ligação máxima, ou seja, considera-se a distância entre os dois elementos mais afastados, um de cada classe. A dissimilaridade entre as funções é quantificada usando duas medidas distintas: a distância euclideana e a distância mínima generalizada (Generalized Minimum distance [12]). Esta última quantifica a "massa" que é necessário mover para transformar uma distribuição na outra. Pode dizer-se que a distância euclideana compara distribuições considerando apenas valores correspondentes, enquanto que a distância mínima generalizada considera a dependência entre valores em diferentes pontos do domínio. Os dendrogramas que resultam da classificação hierárquica são cortados levando à obtenção de uma partição. O nível de corte é decidido com base na análise de dois índices de validação interna da partição, o índice de Calinski-Harabasz [3] e o índice de silhueta [7].

Obtida a partição do conjunto de distribuições, explora-se a similaridade entre os elementos de cada grupo, segundo uma abordagem funcional. Assim, uma distribuição pode ser considerada como atípica por dois motivos: por tomar valores que se afastam do intervalo de valores que a maioria das distribuições apresentam (*outlier* de magnitude); ou por exibir uma forma distinta da maioria (*outlier* de forma). Sobre o conjunto de distribuições de cada grupo é aplicado um método de deteção de outliers que compara não apenas a magnitude das distribuições, mas também a sua forma. O método baseia-se na medida *directional outlyingness* [8]. Este passo foca-se essencialmente nos grupos de menor de dimensão, uma vez que é nestes grupos que se espera detetar distribuições atípicas.
O procedimento pode ser repetido em cada um dos grupos de maior dimensão, para averiguar a existência de sub-grupos que, eventualmente, apresentem padrões distintos. A investigação de níveis mais baixos do dendrograma poderá revelar outros pequenos grupos de interesse.
A metodologia é comparada com os resultados obtidos aplicando apenas o referido método de deteção de *outliers* (segundo passo) a todo o conjunto de dados, tal como descrito em [10].

3. APLICAÇÃO A DISTÂNCIAS GENÓMICAS

O procedimento foi aplicado num conjunto de distribuições de distâncias entre palavras. Focamo-nos em palavras de tamanho 5 e nos seus padrões de distribuição ao longo do genoma humano completo. O conjunto é formado por 1024 distribuições e são consideradas distâncias inferiores a 1000.
A análise de classificação hierárquica do conjunto de dados conduziu à obtenção de cinco classes, com 53 (C1), 162 (C2), 705 (C3), 7 (C4) e 97 (C5) elementos, respetivamente. Do ponto de vista gráfico, as distribuições pertencentes às classes C1, C4 e C5 apresentam uma tendência global semelhante dentro de grupo e bem demarcada da tendência dos outros grupos. De facto, nas duas classes de menor dimensão, C1 e C4, não são identificadas distribuições *outliers*; na classer C5 há dois elementos que são marcados como *outliers*, estando muito próximos do limiar entre *outlier* e não *outlier*. Relativamente à classe C2, são marcados como *outliers* 5 observações, duas das quais se afastam claramente dos restantes elementos do grupo. Por fim, é na classe de maior dimensão, C3, que se encontra uma maior variedade de distribuições, sendo 31 marcadas como *outliers*.
Desta primeira análise resulta a existência de dois grupos de distribuições, com padrões semelhantes entre si, mas que se evidenciam dos restantes padrões (C1 e C4). Sendo a classe C3 muito heterogénea, procedemos a uma análise classificatória neste ramo do dendrograma (*subclustering*). A investigação deste nível mais baixo do dendrograma revelou quatro (sub)classes, duas de maior dimensão (sC1: 291, sC2: 403) e duas de dimensão reduzida (sC3: 7, sC4: 4). Nas duas classes de menor dimensão não foram identificadas distribuições outliers.Pelo que, desta segunda análise, resulta a existência de dois grupos de distribuições homogéneas cujos padrões se demarcam dos restantes (sC3 e sC4).

3. CONCLUSÕES

O objetivo do procedimento assenta na determinação automática de grupos de palavras genómicas com padrões de distribuição similares entre si e afastados dos da maioria das palavras. A ideia chave do nosso procedimento consiste em selecionar distribuições pertencentes a classes de menor dimensão e, nestes, avaliar a similaridade entre as distribuições através de um método de deteção e *outliers* funcional.
A aplicação do procedimento no conjunto de palavras de tamanho 5 permitiu identificar quatro grupos de distribuições bem definidos: cada grupo é formado por distribuições com padrões muito semelhantes, uma vez que nenhuma das suas distribuições é marcada como *outlier*. Dois dos grupos identificados são formados unicamente por distribuições marcadas como *outliers* na análise global, o que sugere que o método proposto deteta grupos de distribuições que se destacam da maioria.

AGRADECIMENTOS

Este trabalho é parcialmente financiado pelo FEDER (Fundo Europeu de Desenvolvimento Regional) e FCT (Fundação Portuguesa para a Ciência e Tecnologia) através dos projetos UID/MAT/04106/2013 do CIDMA (Centro de Investigação e Desenvolvimento em Matemática e Aplicações), UID/EEA/50014/2013 do INESC-TEC (Instituto de Engenharia de Sistemas e Computadores do Porto), POCI-01-0145-FEDER-006961 do COMPETE 2020 (Programa Operacional Competitividade e Internacionalização) e bolsa de doutoramento PD/BD/105729/2014 de AT.

Referências

[1] Afreixo, V., Bastos, C. A. C., Pinho, A. J., Garcia, S. P., Ferreira, P.J. S. G.: Genome analysis with inter-nucleotide distances. Bioinformatics. **25** (23), 3064—3070 (2009).

[2] Burge, C., Campbell, A.M., Karlin, S.: Over-and under-representation of short oligonucleotides in DNA sequences. Proceedings of the National Academy of Sciences **89**(4), 1358–1362 (1992)

[3] Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods **3**(1), 1–27 (1974)

[4] Craven, P., Wahba, G.: Smoothing noisy data with spline functions. Numerische mathematik **31** (4), 377–403 (1978)

[5] MacIsaac, K.D., Fraenkel, E.: Practical strategies for discovering regulatory DNA sequence motifs. PLoS computational biology **2**(4), e36 (2006)

[6] Nakamoto, T.: Evolution and the universality of the mechanism of initiation of protein synthesis. Gene **432**(1), 1–6 (2009)

[7] Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)

[8] Rousseeuw, P. J., Raymaekers, J., Hubert, M.: A measure of directional outlyingness with applications to image data and video. Journal of Computational and Graphical Statistics (just-accepted) (2017)

[9] Silverman, B. W.: Some aspects of the spline smoothing approach to non-parametric regression curve fitting. Journal of the Royal Statistical Society. Series B, 1–52 (1985)

[10] Tavares, A. H. M. P., Afreixo, V., Brito, P., Filzmoser, P.: Directional Outlyingness applied to distances between Genomic Words Proceedings 22nd Portuguese Conference on Pattern Recognition. Aveiro. (2016).

[11] Tavares, A. H. M. P., Pinho, A. J., Silva, R. M., Rodrigues, J. M. O. S., Bastos, C. A. C., Ferreira, P. J. S. G., Afreixo, V.: DNA word analysis based on the distribution of the distances between symmetric words. Scientific Report **7** (728), (2017)

[12] Zhao, X., Sandelin, A.: GMD: measuring the distance between histograms with applications on high-throughput sequencing reads. Bioinformatics **28** (8), (2012)