Leis que governam a estrutura primária do ADN dos seres vivos

Vera Afreixo[1, 2], *vera@ua.pt*
Ana Helena MP Tavares[1], *ahtavares@ua.pt*

1. *Universidade de Aveiro*
2. *iBiMED (instituto de Biomedicina)*

Introdução

Qualquer computador pessoal é capaz de gerar sequências de As, Cs, Gs e Ts com tamanhos iguais aos dos genomas dos seres vivos e até com palavras genómicas com probabilidades iguais às dos genomas reais. No entanto, até um supercomputador, a menos da produção de uma cópia, é incapaz de produzir genomas funcionais. Atualmente ainda se está longe de encontrar modelos capazes de gerar sequências de nucleótidos que sejam funcionais!

O que distancia uma sequência aleatória de nucleótidos de um genoma real? Se assumirmos que os genomas foram criados a partir de uma junção aleatória de nucleótidos esta divergência entre o genoma real e o aleatório poderá ter informação sobre o processo evolutivo das espécies. A descoberta do modelo gerador do genoma de cada espécie poderia trazer grandes contribuições para a caracterização do processo evolutivo das espécies. Mas este é, aparentemente, um problema muito complexo.

É prática comum tentar encontrar a solução para um problema complexo através da decomposição em sub-problemas mais simples.

Neste texto, essencialmente, abordaremos o sub-problema da caracterização das distâncias entre símbolos genómicos e o das extensões da segunda lei de Chargaff avalia a divergência de comportamentos das sequências reais e aleatórias.

Uma sequência simples de ADN pode ser vista como uma sequência de quatro letras A, C, G, T onde A-T e C-G são pares de bases complementares. Uma palavra genómica é um subconjunto de letras que surgem justapostas na sequência de ADN em que o tamanho da palavra é dado pelo número de letras que a compõe.

O complemento invertido de uma palavra genómica é uma palavra que se obtém da palavra inicial alterando cada um dos nucleótidos pelo nucleótido complementar e invertendo a ordem pela qual os nucleótidos se apresentam (e.g. o complemento invertido da palavra ACCGT é ACGGT).

Numa sequência genómica a distância entre nucleótidos pode ser definida como a diferença entre as posições de duas ocorrências sucessivas do mesmo símbolo. Para concretizar esta ideia, considere-se a sequência AACGTCGAAATCCGTAA cujas quatro sequências de distâncias entre nucleótidos são $d^A = (1; 6; 1; 1; 6; 1)$; $d^C = (3; 6; 1)$; $d^G = (3; 7)$; $d^T = (6; 4)$. De modo análogo pode determinar-se a distância entre palavras genómicas (Afreixo *et al.*, 2015a).

O afastamento de uma sequência genómica real a uma sequência gerada num cenário aleatório pode traduzir, de algum modo, a evolução natural das espécies defendida por Darwin. O efeito da aleatoriedade pode ser avaliado através da dissimilaridade entre a sequência real e a sequência gerada num contexto aleatório controlado, em particular, gerada sobre o pressuposto da independência entre símbolos.

Neste artigo será discutida a evolução seletiva dos seres vivos na estrutura primária dos seus genomas em dois aspetos que são usados para caracterizar os genomas das espécies: a relação entre as ocorrência das palavras genómicas complementos invertidos entre si (extensões da segunda lei de Chargaff) e a distribuição de distâncias entre palavras genómicas.

Extensões da segunda lei de Chargaff

As extensões da segunda lei de Chargaff têm vindo a ser referidas como uma lei universal que está presente nos genomas das espécies, sendo este fenómeno também designado por simetria em cadeia simples de ADN (designação adotada). Esta simetria é caracterizada pela semelhança entre as frequências das palavras genómicas e dos correspondentes complementos invertidos. A Fig. 1 apresenta um exemplo relativo à sequência completa do genoma humano para as palavras de tamanho três (também designadas por trinucleótidos). Na figura observa-se que os pares de palavras que são complementos invertidos entre si apresentam frequências de ocorrência próximas e que, de modo geral, apresentam maiores diferenças em relação às frequências de ocorrência das restantes palavras do genoma.

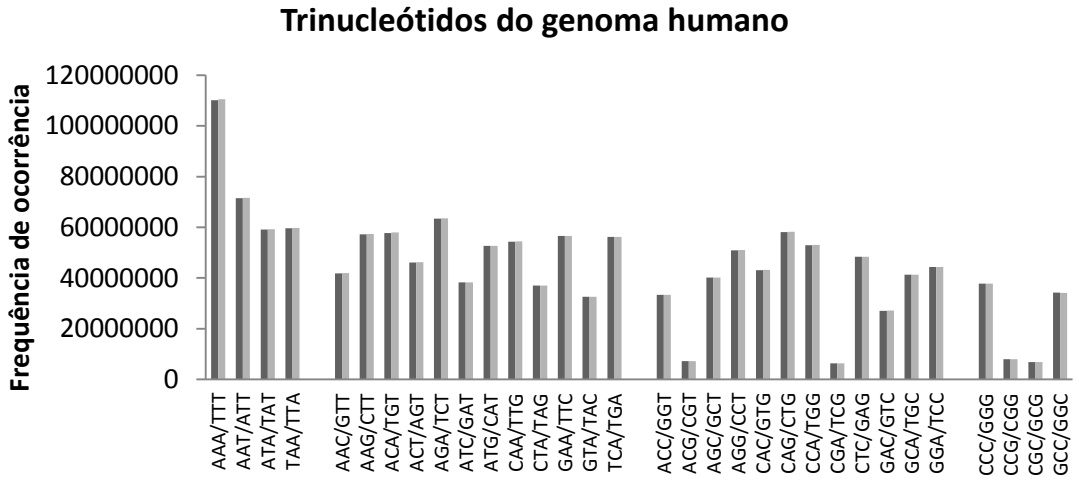

Figura 1: Frequência de ocorrência dos trinucleótidos no genoma humano (versão 37.3).

A avaliação desta regra tem sido feita usando muitas abordagens diferentes, desde a simples correlação entre as frequências das palavras e dos seus complementos invertidos (Baisnée *et al.*, 2002) até abordagens mais exigentes do ponto de vista computacional como a distância de simetria entre palavras baseada na medida de Ulam (Aldous and Diaconis, 1999).
A Tabela 1 apresenta dados relativos à avaliação da simetria em cadeia simples de ADN do genoma humano, para as palavras genómicas até tamanho 10, usando o coeficiente de Pearson, $r \in [-1,1]$, e uma distância baseada na medida de Ulam proposta em Afreixo et al. (2013), $W_s \in [0,1]$.

k	1	2	3	4	5	6	7	8	9	10
r	1	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999
Ws	0	0	0	0,1563	0,3516	0,6763	0,9221	0,9795	0,9967	0,9983
VR	1	210,8	247,4	262,9	270	255,6	214,1	165,3	126,8	103,5

Tabela 1: Avaliação da simetria em cadeia simples de ADN do genoma humano, para as palavras genómicas até tamanho 10, usando o coeficiente de correlação de Pearson (*r*), uma distância baseada na medida de Ulam (W_s) e uma medida de simetria excecional (*VR*).

Os resultados da correlação mostram efeitos de simetria muito fortes, mas estes valores são de certa forma enganadores uma vez que as frequências de ocorrência das palavras constituídas só por As e só por Ts apresentam ordem de grandeza diferente das restantes palavras[1]. A distância W_s é mais robusta

[1] Nas sequências genómicas ocorrem frequentemente poli-As (subsequências de sucessivos As) e poli-Ts (subsequências de sucessivos Ts).

a esta diferença da frequência entre palavras e mostra que para palavras até tamanho quatro o efeito de simetria é forte, mas a partir daí diminui substancialmente. A distância W_s também não é uma medida perfeita de simetria, pois a presença de pequenas variações entre as frequências das diferentes palavras pode contribuir para que esta assuma valores elevados.

Acredita-se que a ocorrência deste tipo de simetria tem motivação biológica como, por exemplo, as estruturas de "stem-loop". Porém, a prevalência de um só fenómeno biológico, por si só, poderá ser insuficiente para explicar a ocorrência do fenómeno de simetria em cadeia simples de ADN. E por outro lado, um modelo aleatório simples poderá gerar sequências que evidenciam um forte efeito de simetria em cadeia simples de ADN, e.g. modelo de independência entre nucleótidos assumindo probabilidades de ocorrência iguais para nucleótidos complementares. Neste caso, constata-se que os grupos de composição equivalente[2] são constituídos por palavras de igual probabilidade e, em particular, as palavras que são complemento invertido entre si têm a mesma probabilidade.
Motivado pela afirmação de Qi *et al.* (2004) que reconhece que a divergência entre o genoma real e o aleatório traduz a evolução seletiva, foi proposta por Afreixo *et al.* (2015) a análise da simetria excecional realçando a semelhança das frequências das palavras genómicas complemento invertidos entre si, com a frequência das palavras genómicas de composição equivalente. A medida de simetria excecional também incorpora conceitos relacionados com testes e medidas de tamanho do efeito de ajustamento, confrontando distribuições empírica e de referência.
A medida de simetria excecional, *VR*, assume valores positivos, o valor 1 não traduz qualquer simetria excecional e valores maiores do que 1 traduzem a simetria excecional (Afreixo *et al.*, 2015b). A Tab.1 apresenta os resultados da medida de simetria excecional, *VR*. Naturalmente, verifica-se que para os nucleótidos não há simetria excecional, enquanto que para as restantes palavras até tamanho 10 existe.
Com esta medida não é possível identificar a origem do fenómeno de simetria em cadeia simples de ADN, mas introduz-se mais conhecimento sobre o fenómeno avaliando, simultaneamente, a semelhança entre as frequências das palavras complemento invertido entre si e a divergência do fenómeno de simetria em estrutura de independência entre nucleótidos. Desta forma através da medida de simetria excecional, acreditamos que estamos a medir a evolução seletiva das palavras no sentido de favorecer mais ou menos o fenómeno da simetria em cadeia simples de ADN.,

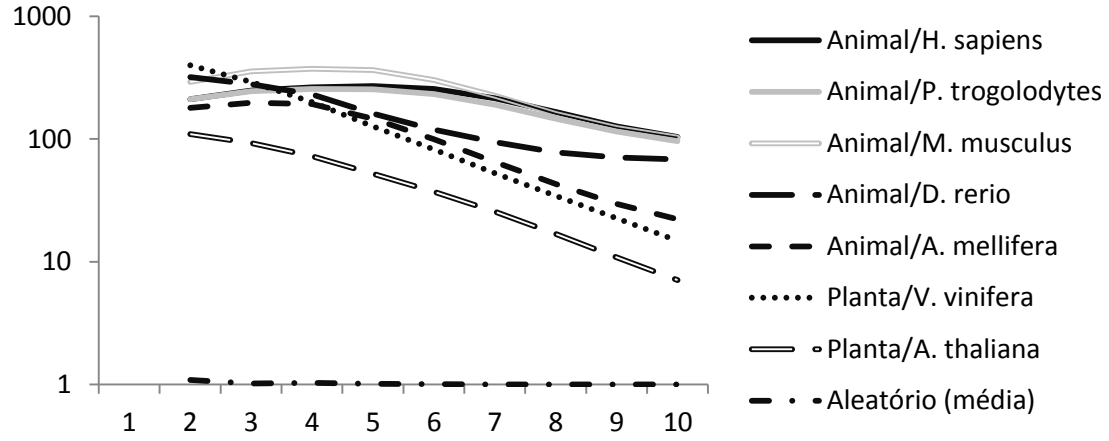

Figura 2: Gráficos de linhas dos valores de simetria excecional para palavras até tamanho 10 (para palavras de tamanho um VR=1). Estão representados seis animais, duas plantas e a média de 100 sequências geradas em contexto de independência, onde as probabilidades dos nucleótidos foram estimadas usando as frequências do genoma humano.

[2] Um grupo de composição equivalente é um conjunto de palavras genómicas que na sua composição têm o mesmo número de A ou T e o mesmo número de C ou G (e.g. as palavras AAC,TCA e TTG pertencem a um mesmo grupo de composição equivalente).

A Fig. 2 apresenta os valores de simetria excecional para diferentes espécies (seis animais incluindo o homo sapiens e duas plantas) e a média dos valores da medida de simetria excecional de 100 sequências aleatórias geradas em contexto de independência de nucleótidos onde a probabilidade de cada nucleótido é estimada pela probabilidade de ocorrência no genoma humano.
Observa-se que há uma clara diferença entre os genomas reais e os aleatórios, sendo que os genomas reais apresentam simetria excecional. Os valores de simetria excecional variam com o tamanho da palavra genómica e apresentam diferenças entre as espécies, de destacar que os gráficos de linhas mais parecidos são os do genoma humano (H. sapiens) e do genoma do chimpanzé (P. trogolodytes).

Distâncias entre palavras genómicas

As distribuições de distâncias entre palavras genómicas têm vindo a ser exploradas em diferentes contextos, tais como a comparação do genoma de diferentes espécies, bem como a distinção entre as diferentes partes constituintes do ADN (e.g. ADN mitocondrial, ilhas de CpG).
A partir da sequência de distâncias, para uma determinada palavra, é então possível definir a distribuição empírica das distâncias entre palavras. No exemplo AACGTCGAAATCCGTAA, e para o caso das palavras AA, as distâncias 1 e 7 têm uma frequência relativa de 1/3 e 2/3, respetivamente.
Uma forma de determinar propriedades estatísticas de diferentes genomas é estudar e caracterizar a distribuição de distâncias entre palavras genómicas. Assumindo que a sequência de palavras é gerada aleatoriamente e que as palavras são independentes e identicamente distribuídas, a distribuição de distâncias entre palavras é a distribuição geométrica.
No entanto, na maioria dos estudos as palavras genómicas são contadas tendo em conta a sobreposição de palavras adjacentes, sendo óbvia a não independência entre as palavras da sequência. Assim, numa abordagem que assuma a sobreposição de palavras, impõe-se a definição de outras distribuições de referência.
A distância entre ocorrências sucessivas de padrões de letras é um assunto estudado, com muitos resultados teóricos já deduzidos, em particular, no que diz respeito à função que traduz o tempo de espera até ao retorno a um padrão específico (e.g. Robin *et al.* (2001)). No caso mais simples em que se pretende obter a distribuição de distâncias entre palavras em sequências cujos nucleótidos são gerados independentemente, e admitindo a sobreposição entre palavras, poderemos recorrer a um diagrama de estados que traduz o progresso na identificação da palavra à medida que cada símbolo é lido na sequência (ver Afreixo *et al.* (2015a)). Sob o pressuposto de independência dos nucleótidos, qualquer transição para um estado está associada à probabilidade de ocorrência do novo nucleótido que é lido na sequência. Note que cada distância está associada ao número de transições necessárias até obter o sucesso (o estado absorvente) e aos diferentes percursos que se podem efetuar até o obter.
Numa sequência genómica real, as distribuições de distâncias entre palavras são diferentes das distribuições associadas a uma sequência gerada aleatoriamente em contexto de independência entre os nucleótidos. Por exemplo, no humano, verifica-se que a distribuição de distâncias associada a uma palavra é muito semelhante à distribuição de distâncias associada ao seu complemento invertido (Tavares *et al.*, 2015) e que a distribuição de distâncias do CG apresenta características muito diferentes das distribuições dos restantes nucleótidos (Afreixo *et al.*, 2015).
A Fig. 3 apresenta as distribuições de distâncias da palavra CG do genoma humano e da distribuição de referência em contexto de independência entre nucleótidos, evidenciando-se algumas diferenças: até cerca da distância 50 as frequências empíricas são inferiores às do modelo de referência e para distâncias maiores são superiores.
A geração de modelos para a distribuição de distâncias entre palavras em contexto aleatório apresenta-se como tópico de estudo de interesse, e acredita-se que permite evidenciar a contribuição da evolução seletiva no ADN das espécies.
De modo a avaliar a divergência de comportamentos entre sequências reais e sequências geradas em cenários aleatórios, deve subtrair-se o efeito da aleatoriedade à sequência real (Qi *et al.*, 2004). A análise de resíduos apresenta algum potencial na construção de árvores filogenéticas. Por exemplo, na comparação entre a distribuição de distâncias real e a distribuição de referência pode recorrer-se ao erro relativo (Afreixo *et al.*, 2009).

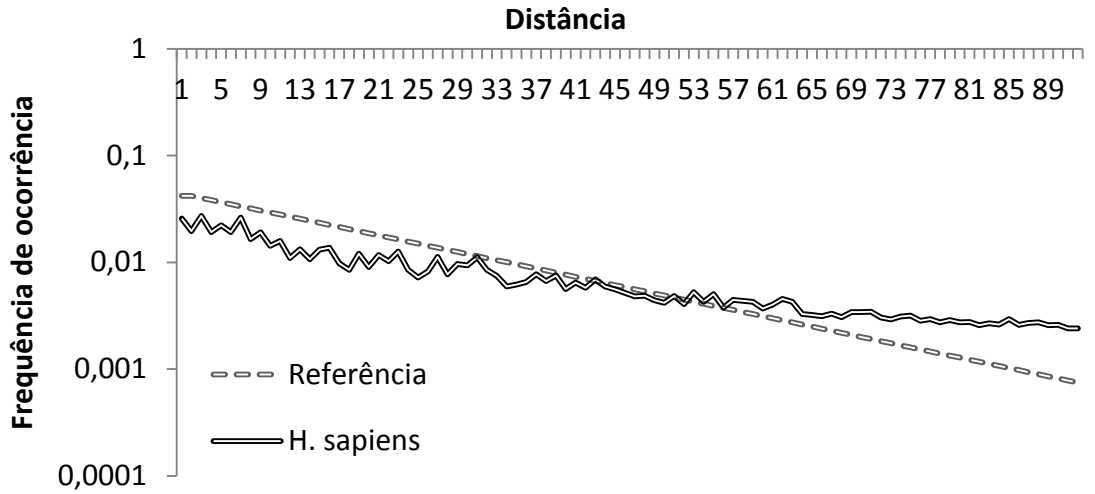

Figura 3: Gráficos de linhas das distribuições de distâncias da palavra CG do genoma humano e da distribuição de referência em contexto de independência entre nucleótidos.

O vetor de erros relativos associado a um genoma pode ser visto como uma assinatura genómica que identifica a espécie, podendo ser usados na construção de dendrogramas muitas vezes interpretados como árvores filogenéticas. As espécies são hierarquicamente agrupadas e as dissimilaridades podem ser vistas como distâncias evolutivas. Por exemplo, se o objetivo for comparar as espécies tendo por base as distribuições de distâncias entre palavras de um determinado tamanho, k, pode definir-se uma distribuição global, bastando para tal aplicar a lei de probabilidade total a todas as 4^k distribuições. O dendrograma da Fig. 4 foi construído com base nas distribuições de distâncias globais entre dinucleótidos, para genomas completos, aplicando o método de ligação média e usando a distância euclidiana no cálculo da matriz de similaridade. O corte em quatro classes evidencia uma separação das espécies em quatro grupos bem distintos: animais, plantas, fungos e bactérias.

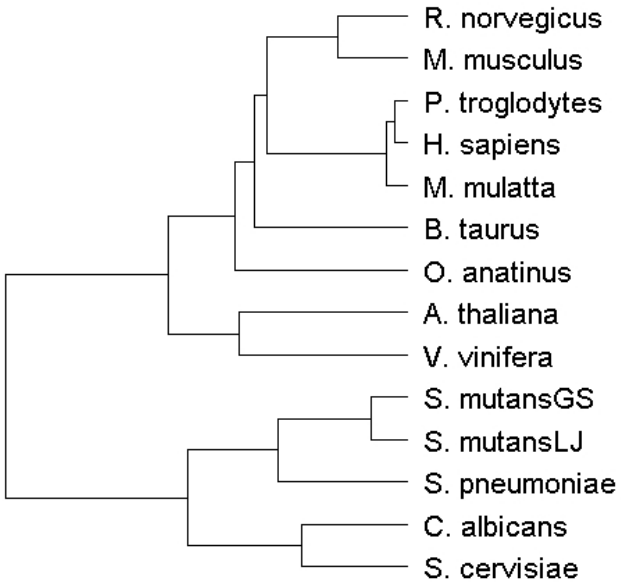

Figura 4: Dendrograma das assinaturas genómicas (definidas a partir dos erros relativos entre as distribuições das distâncias globais das palavras de tamanho 2 e a distribuição de referência global obtida em contexto de independência de nucleótidos). Foram considerados sete animais (R. norvegicus, M. musculus, P. troglodytes, H. sapiens, M. mulatta, B. taurus, O. anatinus), duas plantas (A. thaliana, V. vinífera), dois fungos (C. albicans, S. cervisiae) e três bactérias (S. mutansGS, S. mutansLJ, S. pneumoniae).

Conclusão

A partir dos conceitos de distâncias entre palavras e de simetria em cadeia simples de ADN é possível extrair perfis, diretamente relacionados com as características do ADN, mostrando potencial na discriminação entre espécies e na caracterização de estrutura primária do ADN das espécies.

A explicação da evolução seletiva dos organismos vivos é um tópico que provavelmente não terá uma resposta única e naturalmente não há forma de avaliar a veracidade das conclusões que tenham e venham a ser tiradas. No entanto, poder avaliar o desempenho de um procedimento na análise de dados genómicos tem sido uma experiência fantástica para os investigadores que abraçam estes temas. A criação de assinaturas genómicas poderá ter grande potencial na identificação automática de espécies, principalmente em áreas tão importantes como a da microbiologia.

Referências

Afreixo, V., Bastos, C. A. C., Garcia, S. P., Rodrigues, J. M. O. S., Pinho, A. J., Ferreira, P. J. S. G., 2013. The breakdown of the word symmetry in the human genome. *Journal of Theoretical Biology* 335, 153–159.

Afreixo, V., Bastos, C. A. C., Rodrigues, J. M. O. S., Silva, R. M. 2015a. Identification of DNA CpG islands using inter-dinucleotide distances. *Optimization in the Natural Sciences: Communications in Computer and Information Science* 499, 162–172.

Afreixo, V., Rodrigues, J.M., Bastos, C.A.C., 2015b. Analysis of single-strand exceptional word symmetry in the human genome: new measures. *Biostatistics*, 16, 209–221.

Afreixo, V., Bastos, C. A. C., Pinho, A. J, Garcia, S. P. and Ferreira, Paulo J. S. G., 2009. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 23, 3064–3070.

Aldous, D., Diaconis, P., 1999. Longest increasing subsequences: from patience sorting to the Baik–Deift–Johansson theorem. *Bull. Am. Math. Soc*. 36, 199–213.

Baisnée, P.F., Hampson, S., Baldi, P., 2002. Why are complementary DNA strands symmetric? *Bioinformatics* 18, 1021–1033.

Qi, J., Wang, B., Hao, B.I., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal of Molecular Evolution* 58, 1–11.

Robin, S., Daudin J.J., 2001. Exact distribution of the distance between any occurrences of a set of words. *Annals of the Institute of Statistical Mathematics*, 53 (4), 895–905.

Tavares, A.H., Afreixo, V., Rodrigues, J. M. O. S., Bastos, C. A. C., 2015. The symmetry of oligonucleotide distance distributions in the human genome. *ICPRAM*, Vol. 2 pp. 256-263, Science and Technology Publications.