UNIVERSIDADE FEDERAL DA PARAÍBA

PROGRAMA DE PÓS GRADUAÇÃO EM FÍSICA

MESTRADO EM FÍSICA

Análise de Similaridade de Sequências Genômicas.

Ítallo Costa Fonseca

João Pessoa - PB

2013

Ítallo Costa Fonseca

Análise de Similaridade de Sequências Genômicas.

Dissertação apresentada ao Programa de Pós Graduação em Física da UFPB, como requisito parcial para obtenção do título de mestre em Física.

Orientador: Dr. Edvaldo Nogueira Jr.
Co-Orientador: Dr. Pedro Hugo de Figueirêdo (UFRPE)

João Pessoa - PB

2013

Ítallo Costa Fonseca

Análise de Similaridade de Sequências Genômicas.

Dissertação apresentada ao Programa de Pós Graduação em Física da UFPB, como requisito parcial para obtenção do título de mestre em Física.

BANCA EXAMINADORA:

_____

Prof. Dr. Edvaldo Nogueira Júnior - UFPB

_____

Prof. Dr. Pedro Hugo de Figueirêdo - UFRPE

_____

Prof. Dr. Cláudio B. Silva Furtado - UFPB

_____

Prof. Dr. Sérgio Galvão Coutinho - UFPE

João Pessoa

Agosto de 2013

Dedico este trabalho a minha família.

Agradecimentos

Apesar do curto tempo, mais um ciclo do processo de formação em física chega ao fim. Seria impossível concluir essa etapa e não reconhecer o papel de inúmeras pessoas que contribuiram de forma direta ou indireta para tal. De maneira que, se não cheguei mais longe a culpa foi exclusivamente minha.

Em primeiro lugar, agradeço a Deus pela saúde mental, emocional bem como a oportunidade de conhecer grandes pessoas e ter me conduzido por caminhos de paz.

Agradeço aos meus pais, João e Marinalva, e irmão, Tainer que diante do desafio de ter que ver o filho sair de casa e ir morar em outra cidade em momento algum exitaram e sempre proporcionaram todos os recursos, quando necessários.

Ao meu orientador, Prof. Edvaldo Nogueira Júnior, pela liberdade concedida durante os trabalhos, pelos ensinamentos acadêmicos compartilhados e cultura geral. Seus ensinamentos foram importantes para minha formação e sempre serão lembrados.

Ao meu Co-orientador, Pedro Hugo de Figueirêdo, por desde o início se demonstrar bastante entusiasmado pelo tema, pelas colaborações computacionais necessárias para o desenvolvimento desse trabalho e, também, pelas ideias iniciais que serviram de ponto de partida para nossa abordagem.

Durante minhas primeiras vindas a João Pessoa contei com a ajuda de um grande amigo Thales Lucena no processo de mudança, de forma que, seu apoio foi decisivo para minha permanência.

Quanto a adaptação no departamento de física, agradeço aos amigos de sala Paulo Rogério, Luis Machado e Herondy Mota. Sempre que tive algum problema com conteúdos diversos da física pude contar com cada um deles. Sem deixar de lado os dois baianos que ingressaram juntamente comigo Márcio e Elias.

Aos meus velhos amigos de graduação que estiveram presente em boa parte dessa caminhada Antônio Santos, João Carlos, Priscila Mayana, Abinael, Edinardo Ivison, e Carlos Alberto.

Ao curso de Pós-Graduação em Física da UFPB pelos recursos estruturais e a CAPES que proporcionou a bolsa de mestrado.

Por fim, não menos importante, a minha namorada, Fernanda Freitas, que me conheceu pouco depois do meu ingresso no mestrado e tem me dado toda força e apoio necessário para continuar. Mesmo diante de alguns desânimos, ela soube me reanimar.

"*O macaco é um animal demasiado simpático para que o homem descenda dele*". (Friedrich Nietzsche)

Resumo

Nesta dissertação, investigamos aspectos da similaridade entre sequências completas de DNA mitocondriais. Esta linha de estudo se insere no âmbito da análise de propriedades estatísticas de sequências de DNA baseadas em métodos que buscam entender a informação contida nessas sequências, tema de renovado interesse no contexto dos chamados Sistemas Complexos. Abordagens anteriores foram utilizadas para obtenção das frequências de determinados segmentos de nucleotídeos, considerados como palavras de um dado tamanho, contidos nas sequências. Tais métodos, inspirados em estudos dedicados às propriedades estatísticas de distribuição de palavras em textos linguísticos e sequências simbólicas, podem ser considerados uma alternativa às técnicas e algoritmos de alinhamento de sequências, e têm sido bem sucedidos na descrição de características que permitem inferir similaridade e possíveis critérios de agrupamentos de espécies, ou seja, afinidade biológica entre sequências de DNA. Anteriormente, esta metodologia foi aplicada para avaliar as diferenças entre sequências de DNA codificadas e não codificadas e para extrair aspectos linguísticos dessas sequências através da detecção de palavras-chaves que descrevem informações relevantes embutidas nas sequências. Nesta dissertação, ampliamos tais estudos, no sentido de comparar diretamente o conteúdo de pares de sequências completas de DNA mitocondriais, definindo parâmetros que dependem da distribuição de frequências de palavras das sequências que ressaltam tanto a relevância de determinadas palavras, bem como a possibilidade de agrupamentos de espécies estimando a distância entre essas sequências. Nossos resultados mostram que os melhores agrupamentos entre espécies distintas são obtidos quando calculamos a taxa de aglomeração levando em conta apenas as frequências das palavras. Notamos, também, que quanto maior o tamanho da palavra mais consistente é o agrupamento entre as sequências. A perspectiva de aplicação de nossos resultados,

para analisar também sequências de DNA pertencentes a uma única espécie biológica, pode ser relevante na construção de árvores filogenéticas que são estruturas adequadas para se compreender a história evolucionária dos organismos.

Palavras-chaves: DNA mitocondrial, Frequências de palavras de DNA, Similaridade, Árvores Filogenéticas, Sistemas Complexos.

Abstract

In this thesis, we investigate aspects of similarity between sequences of complete mitochondrial DNA. This line of study falls within the framework from the analysis of statistical properties of DNA sequences based on methods that seek to understand the information contained in these sequences a topic of renewed interest in the context of the so called Complex Systems. Previous approaches were used to obtain the frequencies of certain segments of nucleotides, regarded as the words of a given size, contained in sequences. These methods, inspired by studies devoted to the statistical properties of words distribution in linguistic and symbolical sequences, can be considered an alternative to techniques and algorithms for aligning sequences, and have been successful in the description of characteristics that allow to infer similarity and possible species grouping criteria, it means, biological affinity between DNA sequences. Previously, this methodology has been applied to evaluate the differences between coding and nocoding DNA sequences and to extract linguistic aspects of these sequences by detecting keywords that describe relevant information embedded in the threads. In this dissertation, these studies are expanded in order to directly compare the contents of pairs of complete sequences of mitochondrial DNA, setting parameters that depend on the frequency distribution of sequences of words which highlight both the relevance of certain words as well as the possibility of grouping species estimating the distance between these words. Our results show that the best clusters between different species are obtained when we calculate the rate of agglomeration considering only frequencies of words. We have also observed that the larger the word size is, its greater clustering between sequences. The prospect of applying our results to analyze DNA sequences also belong to a single biological species, may be relevant in the construction of phylogenetic trees that are appropriate structures for understanding the evolutionary history of organisms.

Keywords: mitochondrial DNA, DNA frequencies of words, similarity, phylogenetic trees, complex systems.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Nos últimos anos, as sequências genômicas têm sido objeto de grande atenção por parte dos físicos estatísticos originalmente interessados em correlações de longo alcance caracterizadas por propriedades como leis de potência análogas àquelas presentes em fenômenos críticos e, portanto, no âmbito da matéria condensada. De uma forma geral, no caso das sequências de DNA, um dos desafios da atualidade diz respeito a tentativas de obtenção de leis empíricas a partir do conjunto de informações geradas pelo sequenciamento dos diversos genomas e arquivadas em bancos de dados de acesso público como o **GenBank** [1] e o **PDB** [2], ambos do *Nacional Institute of Health* do governo norte-americano.

A compreensão acerca do funcionamento, papel estrutural, composição química e transmissão de características hereditárias do DNA tem motivado cientistas de diversas áreas desde o início do século XX. Inicialmente, questões como o papel da hereditariedade bem como a descrição da estrutura molecular do DNA necessitaram de notáveis contribuições entre físicos, químicos e biólogos, como por exemplo, Erwin Schrödinger [3], Francis Crick, Linus Pauling, James Watson [4], dentre tantos outros. Portanto, o estudo do DNA desde a sua origem até os dias atuais tem um caráter interdisciplinar.

A primeira identificação da molécula do DNA é atribuída ao bioquímico suíço Johann Friedrich Miescher (Figura 1.1) em 1869, quando o mesmo buscava determinar os componentes químicos do núcleo celular e usava os glóbulos brancos para suas pesquisas. Os

glóbulos brancos eram um bom material, pois são células que apresentam núcleos grandes e fáceis de serem isolados do citoplasma. Em seus estudos, Miescher descobriu um composto de natureza ácida que era desconhecido até o momento. Esse composto era rico em fósforo e em nitrogênio e foi inicialmente denominado por Miescher de nucleína.

Figura 1.1: Johann Friedrich Miescher descobridor da molécula do DNA em 1869 [5].

Após a descoberta de Miescher, deu-se início uma corrida para desvendar, por completo, as principais características dessa macromolécula. Nesse sentido, em 1931, o russo Phoebus Aaron Lavene (1869-1940), estuda a estrutura química dos ácidos nucléicos e identifica seus componentes básicos. Os termos "ácido desoxirribonucléicos"(DNA) e "ácido ribonucléico"(RNA) se tornam de uso comum. Em seguida, no ano de 1943, Oswald Avery descobriu que o DNA é o responsável pela transferência de material genético entre células num processo chamado "transformação". A descoberta sugeria que o DNA seria o material genético básico da célula. Na sequência, em 1949, o austríaco Erwin Chargaff (1905-1992) descobre uma relação quantitativa entre as bases do DNA: a proporção (razão molar) entre adenina e timina e entre guanina e citosina é sempre constante.

Um ano após as contribuições de E. Chargaff, em 1950, Linus Pauling (1901-1994) e Robert Corey (1897-1971) identificaram a estrutura molecular básica de proteínas (o modelo da alfa-hélice). Dois anos depois, eles propõem uma estrutura para o DNA que se mostraria equivocada, com três cadeias helicoidais entrelaçadas (o modelo da tripla hélice). Posteriormente, em 1953, Watson e Crick publicaram a verdadeira estrutura da molécula do DNA: a famosa dupla hélice [6]. Mais adiante, em 1961, Marshall Nirenberg deu início à decifração do código genético e notou que cada trio de bases codificava um aminoácido, "blocos"de construção das proteínas. Apenas em 1977, Richard Roberts e Phil Sharp descobrem, independentemente, que genes de organismos superiores são interrompidos por regiões chamadas íntrons, que não especificam aminoácidos para a formação de proteínas.

Atualmente sabemos que a conformação estrutural de um DNA é do tipo dupla hélice e que toda cadeia do DNA apresenta um conjunto de quatro nucleotídeos: Adenina (A), Citosina (C), Timina (T) e Guanina (G). Normalmente, estes são divididos em dois grupos: purinas (A e G) e pirimidinas (C e T). Outra informação relevante que se tem conhecimento é a de que o DNA armazena mutações que são conservadas com o tempo e, portanto contém informações que preservam a história do organismo. Dessa forma, quando comparamos sequências do DNA podemos inferir a história evolutiva dos organismos, ou seja, a sua filogenia.

Mais recentemente, os estudos estão voltados na direção da caracterização das sequências genômicas segundo uma típica abordagem da linguagem natural, de modo que a aplicação da análise de Zipf e entropia de Shannon são necessárias no intuito de encontrar alguma possível "linguagem genética" [7]. Nessa mesma direção, outros autores suportam a ideia de que partes relevantes das sequências genômicas são detectadas através do conjunto de informações sobre a distribuição espacial, dessas regiões, ao longo da cadeia de DNA [8]. Num passo adiante, outros trabalhos [9] indicam que é possível construir árvores filogenéticas (diagramas que representam as relações de ancestralidade e descendência entre seres vivos) segundo o método de auto atração entre as regiões relevantes da fita. Porém, outros trabalhos suportam a ideia de que não haja conexão entre linguagem natural e uma provável linguagem genética [10]. Em nossa disertação, visamos contribuir para o aprofundamento de tais discussões.

Sob esse aspecto, a dissertação encontra-se organizada do seguinte modo:

No capítulo 2, apresentamos uma visão geral acerca das sequências genômicas, passando pela descrição dos constituintes básicos de uma fita de DNA, pelas propriedades estruturais básicas, descrição do código genético bem como a relação dos 20 aminoácidos e seus respectivos códons, descrição do papel do DNA mitocondrial e as suas características básicas, e, por fim, uma breve descrição do paradoxo da codificação.

Posteriormente, fazemos uma breve revisão, no capítulo 3, sobre os estudos de propriedades estatísticas em sequências de DNA, zelando, sempre que possível, pela ordem cronológica com que os trabalhos foram surgindo. Destacamos alguns trabalho que foram bem sucedidos no processo de deteção de correlações de longo alcance em nucleotídeos bem como o método DFA que leva em consideração a heterogeneidade da sequência, apresentamos, também um método baseado em análise de espectro multifractal que, em última análise propõe a construção de uma árvore filogenética, e, por fim, relatamos brevemente algumas explicações sobre a origem das correlações de longo alcance em sequências de DNA presentes na literatura.

No capítulo 4, apresentamos a metodologia que trata das propriedades estatísticas das denominadas palavras de DNA. Em particular, exibimos os nossos resultados que permitem inferir, estatisticamente, o grau de similaridade entre 60 (sessenta) sequências completas de DNA mitocondrial. As consequências dos resultados obtidos são analisadas e discutidas, como possível critério de agrupamento entre espécies. Finalmente, no capítulo 5, destacamos nossas conclusões e perspectivas de nosso trabalho.

Capítulo 2

Características Gerais do DNA

A busca por unidades fundamentais que determinam as funções responsáveis pela manutenção de uma célula e, consequentemente, da vida recai na estrutura, relativamente simples, do DNA. Dessa forma, o ácido desoxirribonucléico contém as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos. No entanto, tais instruções estão presentes em apenas algumas regiões da cadeia e seu principal papel é armazenar as informações necessárias para a construção das proteínas. A essas regiões, dá-se o nome de codificantes, enquanto que as regiões não codificantes são aquelas que não codificam proteínas e tem um papel, na sequência do DNA, ainda sem resposta.

Outra função atribuída ao DNA é a de transmitir características hereditárias. Com isso, os segmentos dessa molécula que contém a informação genética é denominado de gene. Assim, é esta a unidade fundamental da hereditariedade e cada gene é formado por uma sequência específica de ácidos nucléicos. De maneira simples, os ácidos nucléicos, também chamados de bases nitrogenadas, são os blocos de construção do DNA. Comumente, divide-se esses ácidos em dois grandes grupos: o grupo das purinas correspondendo a adenina e guanina, e, o grupo da pirimidinas correspondendo a citosina e a timina. Vale destacar, que uma base do grupo purina liga-se à apenas uma base do grupo pirimidina e este princípio é conhecido como complementariedade de bases. Na Figura 2.1 apresentamos uma estrutura química típica de cada nucleotídeo.

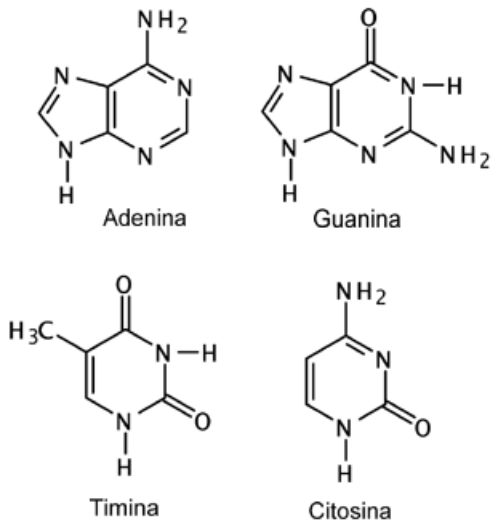

Figura 2.1: Estrutura química dos ácidos nucléicos constituíntes de uma molécula de DNA [11].

Para a formação da molécula de DNA é necessário que ocorra ligação entre os nucleotídeos. Os nucleotídeos estão ligados covalentemente por ligações denominadas fosfodiéster formando entre si pontes de fosfato e, também estão ligados, mais fracamente, a moléculas de açúcares (desoxirribose). Na Figura 2.2 abaixo apresentamos uma representação típica de uma cadeia do DNA.

2.1 Propriedades Estruturais do DNA

No que se refere aos aspectos espaciais, uma cadeia de DNA tem $2,0$ nm a $2,2$ nm de largura, aproximadamente de $3,2$ nm a $3,4$ nm de comprimento e a distância entre dois pares de base consectivos é entre $0,32$ nm e $0,34$ nm (Figura 2.3). Embora os constituíntes fundamentais do DNA possuam dimensões extremamente pequenas, podemos encontar facilmente na natureza cadeias que possuem da ordem de milhões desses constituíntes, como por exemplo, uma sequência típica do *Homo Sapiens* que possui 33.543.332 pares de base.

Em seres vivos, uma molécula de DNA não existe como uma cadeia simples, mas sim como um par de moléculas firmemente associadas. De acordo com o trabalho publicado

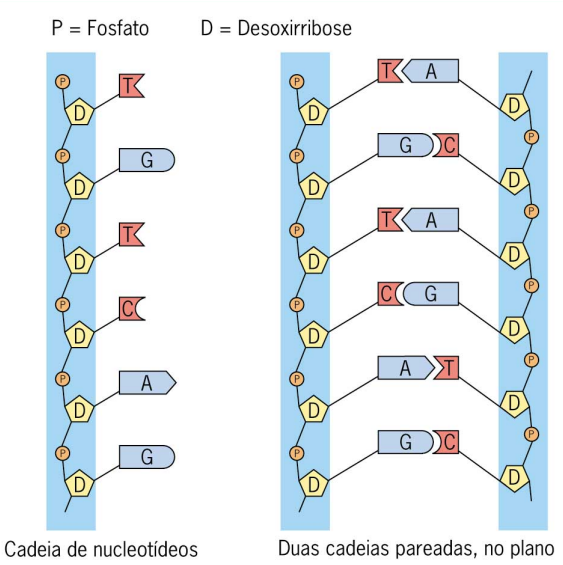

Figura 2.2: Representação de uma cadeia simples e de uma estrutura planar para uma molécula genérica do DNA [12].

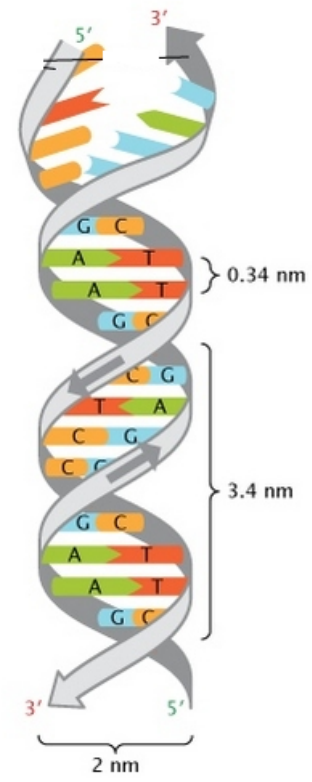

Figura 2.3: Representação espacial de uma molécula de DNA [13].

em 1953, na revista *Nature* (*Molecular Structure of Nucleic Acids*), por J. D. Watson e F. H. C. Crick (ver Figura 2.4), os autores propoem que as duas longas cadeias do DNA enrolam-se como uma "trepadeira" formando uma dupla hélice. Esses autores, elucidaram o modelo tridimensional para a molécula de DNA considerando a natureza química dos seus componentes; as relações de proporcionalidade entre as bases nitrogenadas descritas por Chargaff; e os dados de difração de Raio X obtidos por Rosalin Franklin e Maurice Wilkins. Vale ressaltar que o princípio da complementariedade de bases surgiu nesse contexto. Assim, as purinas combinam-se com as pirimidinas, isto é, A liga-se com T e C com G. Como resultado desta complementariedade, toda a informação contida numa das cadeias de DNA está também contida na outra, o que é fundamental para a replicação do DNA.

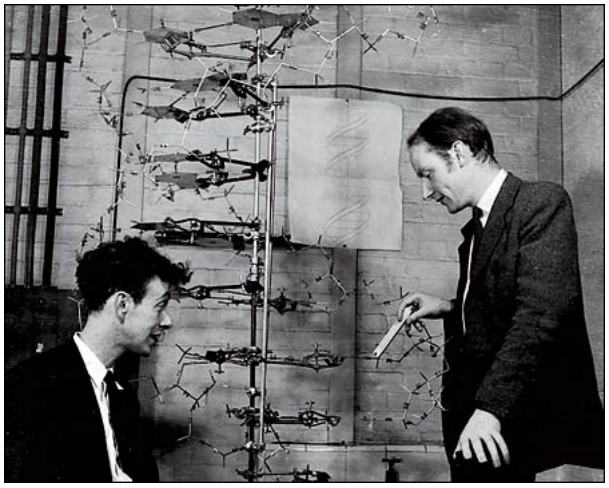

Figura 2.4: A descoberta de Francis Crick (à esquerda) e James Watson (à direita) da estrutura helicoidal da molécula do DNA [14].

Se pensarmos na hélice de DNA como uma estrutura cilíndrica, com cerca de 2 nm de diâmetro, notaremos que a superfície da molécula é irregular, formando dois sulcos ou depressões, de tamanhos diferentes, que giram ao longo de todo o seu comprimento. O sulco menor, resulta da depressão existente entre os giros adjacentes da hélice. Os sulcos são importantes porque deixam livres superfícies para a interação entre o DNA e as proteínas. Na Figura 2.5, apresentamos uma representação estrutural da dupla hélice da molécula do DNA.

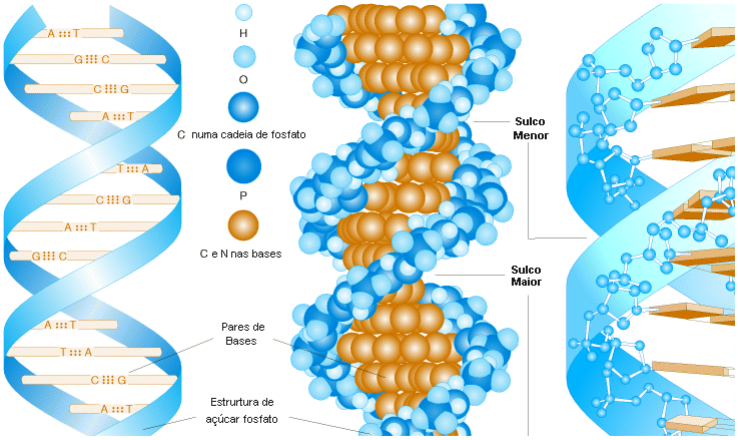

Figura 2.5: Representação esquemática da dupla hélice do DNA [11].

2.2 Código Genético

O código genético é a correspondência entre um determinado tripleto (códons) de nucleotídeos do DNA e um respectivo aminoácido na proteína. A informação genética presente numa molécula de DNA é caracterizada pela ordem e combinação dos quatro nucleotídeos. A leitura da informação é feita através dos códons, ou seja, cada sequência de três bases está associada a um dos 20 diferentes aminoácidos ou a um sinal de início ou de parada da síntese de proteínas. O código genético possui $4^3 = 64$ códons uma vez que, com quatro nucleotídeos só existem 64 combinações possíveis para a construção dos mesmos. Esta associação entre os códons e os aminoácidos correspondentes constitui o código genético.

Noutro aspecto, o código é considerado "degenerado", pois praticamente todos os aminoácidos são determinados por mais de um códon. Um exemplo que descreve bem a "degenerescência "é a glicina (GLY), pois ela é codificada pelos seguintes códons: GGG, GGC, GGA e GGT. Dos vinte aminoácidos existentes apenas a metionina (Met) e o triptofano (Trp) são codificados por um único códon, representados por ATG e TGG, respectivamente. Na Figura 2.6, apresentamos todos os aminoácidos sintetizados por seres vivos bem como seus respectivos códons.

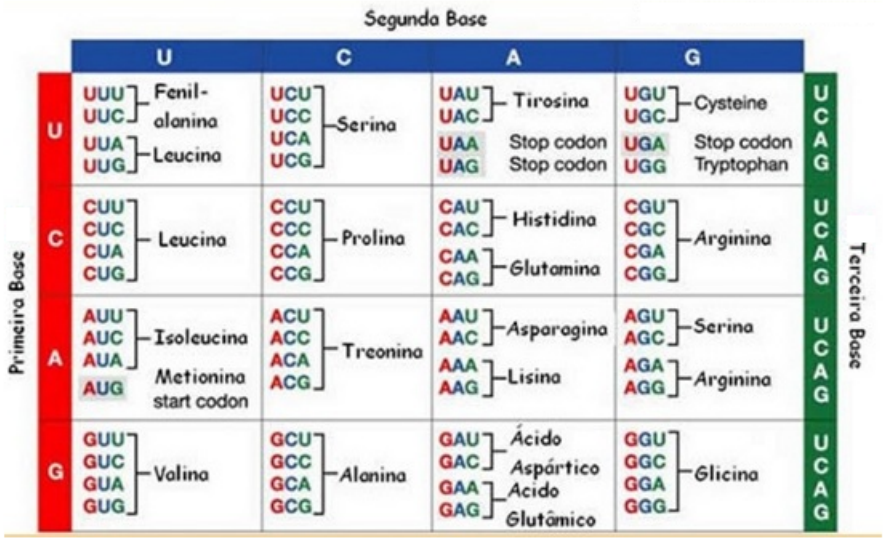

Figura 2.6: Relação dos vinte aminoácidos sintetizados por seres vivos e seus respectivos códons [14].

2.3 DNA Mitocondrial

As mitocôndrias são organelas intracitoplasmáticas envoltas por duas membranas e estão presentes na quase totalidade das células eucariontes cujo núcleo é protegido pelo envoltório nuclear. A principal função atribuída à mitocôndria é a de prover energia à célula. Esta energia é acumulada principlamente em componentes como o trifosfato de adenosina (ATP), que será utilizado quando a célula necessitar de energia para trabalho osmótico, mecânico, elétrico ou químico [16].

Estas organelas têm o seu próprio DNA - que é denominado de DNA mitocondrial (DNAmt) e foi descoberto em 1966 por Van Bruggen, Sinclair e Stevens Nass, mas só foi totalmente sequenciado em 1981, por Anderson e colaboradores [16]. As sequências dos DNAmt apresentam tamanhos típicos da ordem de 10^4 nucleotídeos e, sua cadeia é essencialmente composta por regiões codificantes ($> 60\%$), sendo responsáveis por somente 15% da síntese proteica da cadeia respiratória, o restante é feito pelo DNA nuclear (DNAn).

O DNAmt tem suas próprias características: semi-autônomo, possui um sistema inde-

pendente de replicação, transcrição e translação do seu genoma; herança materna, o DNAmt é herdado da mãe, porque as mitocôndriais presentes nos espermatozóide estão localizadas na cauda deste, que não penetra no óvulo durante a fecundação. Assim, as mitocôndrias presentes no embrião são de herança exclusivamente materna.

2.4 Paradoxo da Codificação

Por um longo tempo, os cientistas certamente esperavam que as variações na anatomia entre os animais fossem refletidas por diferenças claras no conteúdo de seus genomas. Quando comparamos genomas de mamíferos como o camudongo, o rato, o cachorro, o homem e o chimpanzé, no entanto, vemos que seus respectivos grupos de genes são notavelmente similares. O número aproximado de genes no genoma de cada animal (cerca de 20 mil) e as posições relativas de muitos genes se mantiveram praticamente constantes em 100 milhões de anos de evolução. Isso não quer dizer, que não há diferenças no número e na localização dos genes. Mas, à primeira vista, nada nesses inventários gênicos diz "camudongo " ou "cão" ou "humano".

Quando os biólogos avaliam individualmente os genes, de forma detalhada, a semelhança entre espécies também é a regra. As sequências de DNA de duas versões quaisquer de um gene, bem como as proteínas que codificam, são geralmente semelhantes em um grau que simplesmente reflete a quantidade relativa de tempo que se passou desde que as duas espécies divergiram de um ancestral comum. A preservação das sequências codificantes ao longo da evolução é particularmente notória quando consideramos os genes envolvidos na construção e definição das formas do corpo.

Em suma, o parodoxo da codificação, também conhecido por paradoxo C, poderia ser sintetizado pela constatação de que quando comparando-se diferentes espécies, uma com as outras, verifica-se que o tamanho completo dos genomas não apresenta nenhuma correlação com a complexidade fenotípica (características observáveis num organismo) das espécies. Por exemplo, analisando genomas completos de mamíferos e de bactérias nota-se, que na maioria dos casos, esses últimos são significativamente maior que os primeiros.

Capítulo 3

Propriedades Estatísticas de Sequências de DNA

3.1 Correlações em Sequências de Nucleotídeos

Nas três últimas décadas, um dos desafios motivadores, para cientistas de diversas áreas, diz respeito à determinação de aspectos universais da informação codificada no DNA. O processo de caracterização de uma determinada sequência pode servir de ponto de partida para o entendimento de doenças genéticas, tais como, doenças auditivas relacionadas à alterações no DNA mitocondrial, neuropatia (doença do sistema nervoso) óptica, dentre outras [15], fatores de envelhecimento relacionados a mutações nas cadeias do DNA mitocondrial [17], ou, até mesmo, em problemas mais recentes da astrobiologia como a detecção de condições extremas nas quais sequências genômicas possam sobreviver em exoplanetas [18]. Nesse sentido, apresentaremos algumas das técnicas, propostas por diversos autores, que visam encontrar leis empíricas para melhor compreender a sequência do DNA.

Diversos fenômenos naturais não são processos aleatórios independentes, geralmente eles apresentam significantes correlações de longo alcance. Essa característica pode ser notada em vários fenômenos, como por exemplo, em avalanches [19], [20], terremotos [21], extinção de espécies [22], dentre outros. Reconhecer a abrangência de correlações de longo alcance do tipo

lei de potência pode ajudar nos esforços de compreender a natureza, pois uma vez que são encontradas tais correlações, pode-se então quantificá-las segundo um expoente característico. A quantificação deste tipo de comportamento de escala para sistemas aparentemente não correlacionados permite o reconhecimento entre diferentes sistemas, levando, assim, a unificações que poderiam passar despercebidas.

Segundo esse panorama, no estudo de propriedades estatísticas em sequências de DNA, uma técnica que tem apresentado grande aceitação no processo de caracterização de tendências, nas longas cadeias, foi desenvolvido por Peng e colaboradores [23]. O método, trata-se basicamente de um mapeamento unidimensional de uma sequência de nucleotídeos numa caminhada aleatória comumente denominada de (*DNA walk*).

Para o modelo de uma caminhada aleatória unidimensional, um caminhante se move para "esquerda "$[u(i) = -1]$ ou para "direita"$[u(i) = +1]$ de uma unidade de comprimento para cada passo i da caminhada. Quando uma caminhada aleatória for não correlacionada, a direção de cada passo é independente dos passos anteriores, enquanto que numa caminhada correlacionada, a direção de cada passo depende da história ("memória") do caminhante.

Uma vez definida a caminhada aleatória segue-se que a projeção desse modelo para uma fita do DNA obedece às mesmas regras para o "movimento". Assim, o caminhante vai para esquerda ($u(i) = -1$) caso encontre uma purina (A ou G), enquanto que ele vai para a direita ($u(i) = +1$) caso encontre uma pirimidina (T ou C). Sendo assim, o passeio unidimensional permite visualizar diretamente as flutuações entre purinas e pirimidinas de maneira que, inclinações positivas correspondem a uma elevada concentração de pirimidinas, enquanto que inclinações negativas indicam uma maior concentração de purinas.

Feito isso, determina-se uma quantidade estatística que caracteriza a caminhada que é a raiz quadrada média da flutuação $F(l)$, onde l é o tamanho do passo. O comportamento desta grandeza é descrito por uma lei de potência, $F(l) \sim l^{\alpha}$. O caso em que $\alpha = 1/2$ representa a ausência de correlações de longo alcance. Com isso, o método em questão [23] foi aplicado a sequências de DNA codificantes bem como a não codificantes. Os resultados obtidos são que: sequências não codificantes apresentam correlações de longo alcance do tipo lei de potência ($\alpha > 1/2$), enquanto que as sequências codificantes apresentam correlações

de curto alcance ($\alpha = 1/2$).

3.2 Análise de Flutuações sem Tendências

O método anteriormente apresentado foi posteriormente questionado devido ao fato de que, além das flutuações estatísticas normais esperadas em análises de curtas sequências, as regiões codificantes consistem tipicamente apenas de algumas longas regiões, tais regiões levam a alternâncias nas tendências da fita, e por isso, temos uma série que não é estacionária (séries onde a média e o desvio padrão não são localmente constantes). Assim, as análises de escala convencionais não podem ser aplicadas com segurança a toda sequência, mas apenas a subsequências.

A fim de evitar esse problema, Peng e colaboradores [24] desenvolveram um método adequado adaptado especificamente para lidar com problemas associados a séries não estacionárias que é a análise de flutuações sem tendências - o DFA (do inglês *detrended fluctuation analysis*). Esse método se propõe a ser mais eficiente do que o anterior, pois evita falsas deteções de aparentes correlações de longo alcance.

Para implementação do método DFA, inicialmente consideramos a mesma caminhada aleatória segundo as mesmas regras já descritas anteriormente. A mesma quantidade estatística é computada: a raiz quadrada média da flutuação local sem tendências $F_d^2(l)$. Nesse contexto, as sequências de nucleotídeos são tratadas como uma série temporal.

É sabido que a variância apresentará uma dependência com o tamanho da caixa, originando assim, propriedades de escala das flutuações. Dessa forma, se apenas correlações de curto alcance (ou nenhuma correlação) existirem na sequência de nucleotídeos, então a caminhada sem tendências no DNA deverá ter propriedades estatísticas de um passeio aleatório, de forma que $F_d(l) \sim l^{\alpha}$, com $\alpha = 1/2$. Um fato que merece destaque é que se $\alpha < 1/2$, então existirá correlações de longo alcance de modo que teremos uma alternância de anticorrelação (anti-persistência) entre os diferentes tipos de nucleotídeos. Por outro lado, se $\alpha > 1/2$ a correlação também será de longo alcance, no entanto, haverá uma persistência do mesmo tipo de nucleotídeo na sequência.

Mais uma vez, a aplicação desse método revelou que sequências de DNA não codificantes apresentam correlações de longo alcance do tipo lei de potência, enquanto que sequências de DNA codificantes apresentam correlações de curto alcance ou nenhuma correlação.

Diante desse cenário, a aplicação do DFA para a caracterização de propriedades fractais e de correlações de longo alcance em sequências de DNA provou ser eficiente, pois leva em consideração o caráter heterogêneo do DNA que consiste de muitos fragmentos, onde cada um deles apresenta tamanhos diferentes entrelaçados.

3.3 Espectros Multifractais

Numa outra abordagem, J. A. Glazier e colaboradores [9] propõem o mapeamento de uma sequência de DNA segundo uma caminhada aleatória num espaço de quatro dimensões e, estima as correlações de longo alcance presentes em sequências de DNA mitocondriais usando análise multifractal. Num sistema multifractal, o comportamento da função de flutuação em torno de um dado ponto é descrito por uma lei de potência local com distintos expoentes. Nesse sentido, o autor, destaca que o método apresenta uma significativa diferença entre o espectro multifractal de vertebrados e invertebrados.

Em seguida, J. Glazier, utiliza um algoritmo de agrupamento que trata os objetos como pontos num espaço bidimensional, associados aos expoentes de flutuação locais. Finalmente, o algoritmo calcula a diferença entre *clusters* (aglomerados) como distâncias euclidianas no espaço citado a fim de distinguir bem os dois grupos (vertebrados e invertebrados). À medida que a distância entre as espécies se tornam mínimas, temos a associação de que essas espécies são "similares". Espécies diferentes apresentam distâncias maiores.

Dessa maneira, o conjunto de distâncias retiradas do espaço bidimensional descrito acima, permite, *a priori*, reconstruir a história evolutiva dos organismos. A construção das árvores filogenéticas (árvores que mostram a relação evolutiva entre grupos de organismos) revela que além da distinção clara entre grupos de vertebrados e invertebrados, num primeiro instante, um afastamento claro entre mamíferos, anfíbios e peixes. Assim, com a aplicação desse método, temos as primeiras tentativas de deduzir relações biologicamente significativas

a partir de propriedades estatísticas.

Contudo, embora o método de Glazier tenha apresentado resultados aparentemente satisfatórios quanto ao processo de construção de uma árvore filogenética, bem como uma clara distinção entre grupos de invertebrados e vertebrados, a aplicação do método revela-se inconsistente uma vez que trabalhos posteriores demonstraram que sequências de DNA não são multifractais [25]. Sendo assim, se quisermos mapear o processo evolutivo das espécies de maneira mais adequada, outra técnica precisa ser empregada.

3.4 Origem das Correlações de Longo Alcance

Os mecanismos de mutações podem ser um dos fatores responsáveis pela existência de correlações de longo alcance em diferentes escalas de comprimento. As correlações mais longas, nas escalas de comprimento de isocoros (grandes regiões do genoma contendo semelhanças locais quanto à composição de bases) podem se originar devido a mutações de substituição de base durante a replicação. O fato é que diferentes partes dos cromossomos replicam em diferentes fases da divisão celular. Regiões ricas em $C + G$ replicam antes daquelas ricas em $A + T$. Assim a probabilidade de substituição A/T por C/G é maior nas regiões dos cromossomos que se replicam primeiro. Essas taxas de desigualdade nas mutações podem acarretar a formação de isocoros [26].

Quanto às correlações nas escalas de comprimento intermediárias de milhares de pares de base, sua explicação pode ter origem no embaralhamento do DNA por inserção, deleção ou devido a um processo de mutação duplo proposto por W. Li. Por outro lado, as correlações nas escalas de comprimento de várias centenas de pares de base podem ser resultado da simples expansão de repetição. As distribuições de repetições simples são diferentes em entre as sequências codificantes e não codificantes [27].

A distribuição tipo lei de potência nas repetições simples pode ser explicada se assumirmos um processo aleatório multiplicativo para a mutação do comprimento de repetição, isto é, cada mutação leva a uma alteração do comprimento de repetição por um fator aleatório com certa distribuição. Tal processo pode conduzir a erros na replicação ou nos diferentes

cruzamentos. Expansões de repetições simples nas regiões codificantes levariam a uma perda de funcionalidade das proteínas e consequentemente a extinção dos organismos [27].

Assim, a "fraqueza" das correlações de longo alcance no DNA codificante está provavelmente relacionada com a conservação do mesmo durante a evolução biológica. De fato, as proteínas das bactérias e dos seres humanos têm muitos modelos comuns, enquanto que as regiões não codificantes podem ser totalmente diferentes, mesmo para espécies relacionadas. A conservação de proteínas em sequências codificantes e a fraca correlação nas sequências de aminoácidos são, provavelmente, relacionadas com o problema do enovelamento proteico [28].

Capítulo 4

Análise de Similaridade de Sequências Mitocondriais

4.1 Lei de Zipf em Sequências de DNA

Numa tentativa de entender melhor o papel das regiões não codificantes Mantegna e colaboradores [7] aplicaram técnicas de caracterização das linguagens naturais para o DNA. Nesta abordagem, a sequência genômica de uma determinada espécie é interpretada como um texto cujas palavras possuem um tamanho n, onde n é um parâmetro livre. Segundo esses autores as regiões codificantes são constituídas pelos 64 possíveis tripletos reponsáveis pela codificação de aminoácidos, AAA, AAT, ..., GGG. Enquanto que, as regiões não codificantes não apresentam um tamanho característico de palavra.

Diante disso, o método de Mantegna e colaboradores propõe uma extensão da abordagem de Zipf para análise de textos em sequências de DNA codificantes e não codificantes. A análise de Zipf é uma característica presente na linguagem [29]. Essa lei verifica que a frequência f de cada palavra de um texto, e sua categoria (rank) r (a mais frequente recebe categoria 1, a segunda mais frequente categoria 2, e assim por diante), estão relacionadas segundo uma lei de potência $f \sim r^{-k}$, com $k \approx 1$ para todas as linguagens estudadas [29].

Posteriormente, o autor aplica a lei de Zipf para um conjunto de 40 sequências e nota que todas elas apresentam comportamento tipo lei de potência. Porém, o expoente Zipf para as regiões não codificantes apresentam valores maiores do que o das regiões codificantes. Assim, de posse desses resultados, o método suporta a ideia da possibilidade de existência de uma (ou mais do que uma) linguagem biológica estruturada presentes em sequências de DNA não codificantes.

Num trabalho posterior, Mantegna e colaboradores [30] analisam detalhadamente uma segunda característica comum às línguas naturais que é a redundância: letras ou mesmo palavras inteiras que podem ser omitidas ou alteradas sem tornar o texto indecifrável. A noção de redundância foi apresentada e quantificada na clássica obra de Shannon [31], que introduziu o conceito de entropia associada a redundância, de tal modo que existe uma relação explícita entre as duas grandezas. A redundância pode também ser vista como uma manifestação da flexibilidade do código em análise.

A entropia medida para palavras de tamanho n é uma informação parcial sobre o grau de complexidade de uma sequência simbólica [33]. Sendo assim, de forma geral, a entropia para palavras de tamanho n numa dada sequência é definida por $H(n) = -\Sigma_{i=1}^{\lambda^n} p_i log_2 p_i$, onde p_i é a probabilidade de uma palavra i de tamanho n e λ é o número de letras do alfabeto, no caso do DNA $\lambda = 4$. A partir da entropia pode-se calcular a redundância R presente nos textos. A definição da redundância em termos da entropia é descrita pela relação: $R \equiv 1 - \lim_{n\to\infty} H(n)/kn$, onde $k \equiv log_2\lambda$.

Os valores para a redundância indicam que ela aumenta a medida que o tamanho da palavra aumenta. A menor palavra utilizada por Mantegna e colaboradores é $n = 1$, ou seja, palavras como A, C, T e G, ao passo que as maiores palavras apresenta tamanho $n = 7$, como por exemplo, AAAAAAA, AAACTCT, AAAGCGT, ..., GGGGGGG. Outra conclusão é a de que não importa para qual grupo se observe, seja mamíferos, invertebrados ou vírus, todos apresentam um comportamento crescente para a redundância à medida que o tamanho da palavra também aumenta. Muito embora o comportamento da redundância *versus* n seja crescente para os diferentes grupos, ela é significativamente maior para os invertebrados. O terceiro ponto é quando se observa as regiões codificantes, nota-se os menores valores para a redundância, enquanto que para regiões não codificantes, os valores maiores. Por fim, as

sequências completas apresentam valores intermediários para a redundância.

Os resultados descritos até então - para a redundância - levantam a hipótese da existência de uma "linguagem genética" nas regiões não codificantes e que podem fornecer um quadro interpretativo diferente para o paradoxo C. O paradoxo C é um paradoxo que trata do tamanho total do genoma, visto que o tamanho completo dos genomas não parece estar correlacionado com a complexidade fenotípica (características observáveis de um organismo, por exemplo, fisiologia, comportamento, etc.) das espécies. Por exemplo, o tamanho do genoma completo para a espécie *Homo Sapiens* é muito menor do que o da *Ameba Dubia*.

Dessa forma, uma vez destacado o fato de que a informação biológica estruturada seja também armazenada nas regiões não codificantes, pode-se conjecturar que o tamanho global destas regiões deve estar relacionado com a complexidade fenotípica do organismo. Por exemplo, as regiões não codificantes de vertebrados são mais longas do que em invertebrados e têm menor redundância. Isto poderia significar que a quantidade de informações concentrada nas regiões não codificantes dos vertebrados seja maior do que nos invertebrados.

4.2 Detecção de Palavras Relevantes no DNA

Em trabalhos recentes, no contexto de propriedades estatísticas das linguagens naturais, Ortuño e colaboradores [8] desenvolveram um método a fim de detectar palavras chave em textos. Nessa direção, esses autores sugerem que a distribuição espacial das palavras, independentemente da sua frequência relativa, quantifica a sua relevância no texto considerado.

O método consiste na determinação da distribuição $p(x)$ das distâncias de ocorrências sucessivas de uma dada palavra. Alternativamente, também é possível definir a distribuição acumulada $P_1(x) = \Sigma_{x'}^{x} p(x^{'})$. Ambas as funções contém a mesma informação, mas a última é mais útil em termos de precisão numérica.

Como discutido por Ortuño e colaboradores [8] a motivação desse método é baseada no estudo de níveis estatísticos do espectro de sistemas quânticos desordenados, de acordo com a teoria de matrizes aleatórias [32]. Nesse caso, também se faz necessário o estudo

das propriedades de $p(x)$ (ou $P_1(x)$), onde x agora é o espaçamento entre níveis de energia consecutivos. Na Figura 4.1, ilustramos um exemplo típico dos segmentos de um espectro complexo, cada um contendo 50 níveis e todos redimensionados para a mesma escala de modo que a média do espaçamento é unitária em todos os casos.

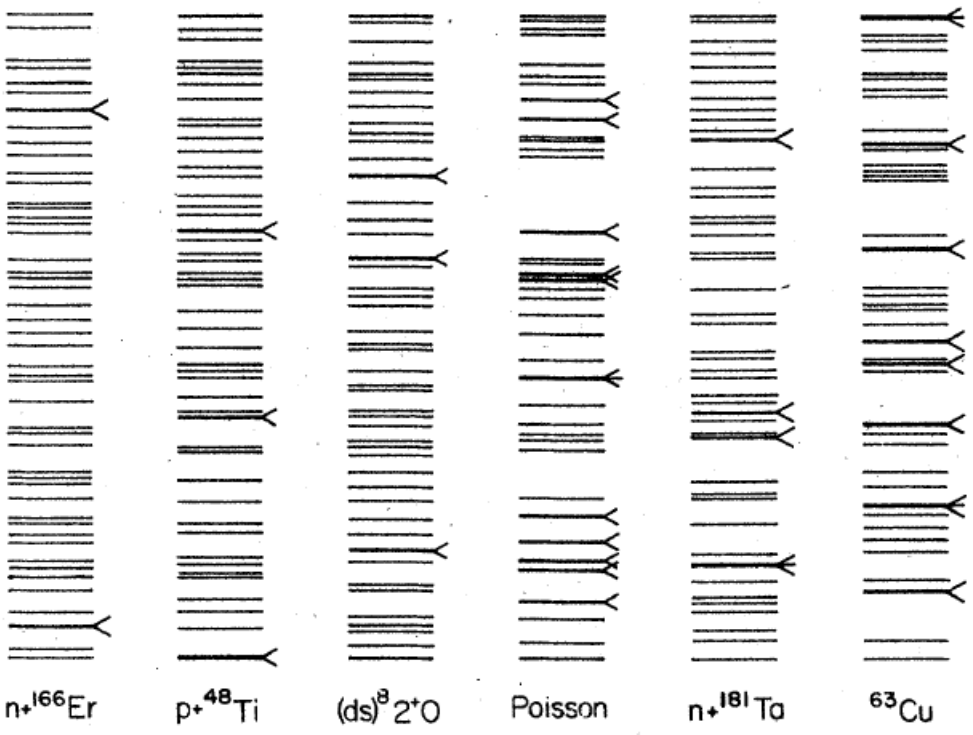

Figura 4.1: Segmentos de espectros complexos, cada um contendo 50 níveis e redimensionados para o mesmo espectro de amplitude de modo que as distâncias médias entre os níveis estão normalizadas [32].

Segundo esta abordagem, qualquer uma das ocorrências de uma palavra em particular é considerada como um "nível de energia" e_i dentro de um "espectro de energia", formado por todas as ocorrências da palavra analisada dentro do texto. O valor de qualquer nível de energia e_i é dado simplesmente pela posição da palavra. No caso de uma palavra relevante os níveis de energia se atraem mutuamente, enquanto que para palavras não relevantes, os níveis de energia não estão correlacionados e, portanto, distribuídos de forma aleatória. Desta forma, quanto maior for a relevância de uma palavra, maior o agrupamento (atração) e maior será o desvio de $p(x)$ em relação ao caso aleatório.

A conexão entre a formação de aglomerados (atração de palavras) e a relevância decorre

do fato de que uma palavra relevante é geralmente o principal assunto em contextos locais e, portanto, ela aparece com mais frequência em algumas áreas e, menos frequentemente em outras, dando origem aos aglomerados (*clusters*). Por exemplo, num livro de mecânica quântica a palavra comutador é uma palavra chave relevante. Ela irá aparecer muitas vezes no contexto dos fundamentos da mecânica quântica e será relativamente rara em outros lugares.

De modo a quantificar a aglomeração (e, assim, a relevância) de uma palavra usando apenas um único parâmetro ao invés de toda distribuição $p(x)$, o que é mais custoso computacionalmente, define-se o parâmetro σ como sendo:

$$\sigma \equiv \frac{s}{\langle x \rangle}, \tag{4.1}$$

com $\langle x \rangle$ sendo a distância média e $s = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$ o desvio padrão de $p(x)$. Para uma palavra particular, σ é o desvio relativo do seu conjunto normalizado de distâncias $\{x_1/\langle x \rangle, x_2/\langle x \rangle, ..., x_n/\langle x \rangle\}$, isto é, distâncias calculadas em unidades da distância média, o que permite a comparação direta dos valores de σ obtidos para palavras com diferentes frequências e torna o parâmetro independente do tamanho do texto analisado.

No contexto da estatística dos espectros de sistemas quânticos desordenados, os níveis de energia não são correlacionados, de maneira que a correspondente distribuição $p(x)$ é a distribuição de Poisson: $p(x) = e^{-x}$, para a qual $\sigma = 1$. Assim, consistentemente, quando a mesma metodologia foi aplicada por Ortuño e colaboradores em textos literários, o valor encontrado para uma palavra sem relevância, sem aglomeração e distribuída aleatoriamente correspondia a $\sigma = 1$, enquanto que os maiores valores de σ correspondia a palavras mais relevantes e que formavam aglomerados.

4.3 Aspectos Linguísticos das Sequências de DNA Mitocondriais

O passo seguinte dado por Ortuño e colaboradores [8] foi estender o método de técnicas estatísticas de coleta de palavras chave em textos literários para as cadeias de DNA, a fim de testar se existem semelhanças entre a linguagem natural e uma possível "linguagem genética". À semelhança dos métodos anteriores, o método em questão interpreta a cadeia de DNA como uma sequência construída a partir de um alfabeto de quatro símbolos, correspondendo aos diferentes nucleotídeos.

O processo de "leitura" do DNA tem como objetivo procurar uma "estrutura de palavra" e é realizado dividindo a sequência do DNA em caixas não sobrepostas de comprimento fixo n (com $n = 1, 2, 3, ...$), associando a cada caixa uma palavra, mudando n vezes a posição inicial de busca a fim de levar em consideração todas as possíveis fases. Feito isso, calcula-se o valor de σ de cada palavra das sequências de DNA utilizadas.

De posse de quatro sequências, sendo duas codificantes (*E. coli* e *S. cerevisiae*) e duas não codificantes (*H. Sapiens* e *C. elegans*), Ortuño e colaboradores analisaram o gráfico de σ *versus* o tamanho n da palavra e verificaram que as sequências codificantes apresentam maiores valores de σ para $n = 3$. Segundo os autores, isso é uma constatação de que o método detecta o "comprimento típico" do código. Por outro lado, quando a sequência estudada é não codificante, os resultados mostram uma atração maior de palavras para os comprimentos de $n = 1$ e $n = 2$ (ver Figura 4.2 adaptada). Os resultados para $n = 1$, segundo Ortuño e colaboradores, são consistentes com as correlações de longo alcance em DNA não codificante previamente relatado por Peng e colaboradores [24] quando se considera mononucleotídeos.

Sob esse panorama, destacamos que os métodos descritos, até então, apontam na direção de que deve existir, de fato, uma relação inerente entre linguagem natural e a sequência do DNA. No entanto, num trabalho proposto por Tsonis e colaboradores [10], utilizando análise de dados e simulações computacionais, chega-se à conclusão de que as sequências de DNA não apresentam propriedades linguísticas.

Em nosso trabalho, visamos aprofundar mais essa discussão no sentido de investigar se o

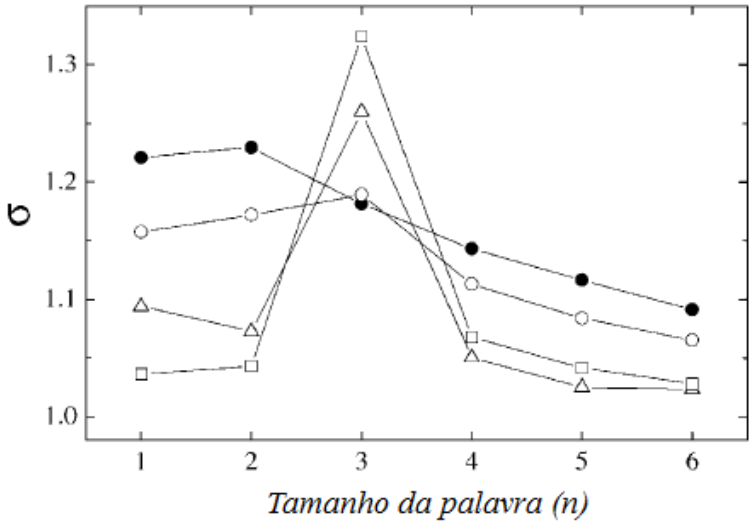

Figura 4.2: Gráfico de σ como função do tamanho da palavra. As duas sequências que apresentam maiores valores de σ para $n = 3$ correspondem às sequências codificantes, enquanto que as outras duas a não codificantes [8].

DNA possui ou não uma "linguagem genética" com as mesmas características da linguagem natural. Para tanto, diferentemente dos trabalhos anteriores que utilizaram 15 [9], 37 [7], 25 [30], 36 [33], 4 [8] e 2 [34] sequências, coletamos no GenBank [1] um total de 60 (sessenta) sequências mitocondriais, para obter, assim, uma maior robustez estatística. Embora o DNA mitocondrial seja mais curto que o DNA nuclear, ele é suficientemente grande para análise estatística em muitos animais. Além do que, os métodos de análise são mais fáceis de aplicar a tais sequências cujos comprimentos são da ordem de 10^4 pares de base (pb), enquanto que o DNA nuclear tem da ordem de 10^6 de pares de base.

Feito isso, classificamos nossas sequências em quatro grupos: o grupo 1 contém 15 sequências de invertebrados; o grupo 2 contém 15 sequências, dentre os quais temos répteis (9), anfíbios (3) e peixes (3); o terceiro grupo apresenta 15 sequências de aves, e, por fim, o grupo 4 contendo apenas sequências de mamíferos. Uma descrição das espécies utilizadas bem como códigos de acesso no GenBank e o tamanho das sequências dos grupos 1 e 2 se encontram na Tabela 4.1, enquanto dos grupos 3 e 4 estão na Tabela 4.2, respectivamente.

Durante a "leitura" da sequência, utilizamos a mesma metodologia empregada por Ortuño e colaboradores. Porém, em nosso método, o tamanho mínimo para uma palavra

é $n = 1$, enquanto que o máximo é $n = 7$. Nesse processo, escolhida uma dada palavra, computamos as distâncias entre as ocorrências sucessivas, em unidades de nucleotídeos. Na Figura 4.3, vê-se uma representação esquemática para o cálculo da distância entre ocorrências sucessivas de uma palavra específica TG presente num pequeno fragmento da sequência do *Homo Sapiens*. Notamos que, para esse exemplo, a distância média é $\langle x \rangle = 16,2$, a distância quadrática média é $\langle x^2 \rangle = 757$ e o desvio relativo $\sigma = 1,37$.

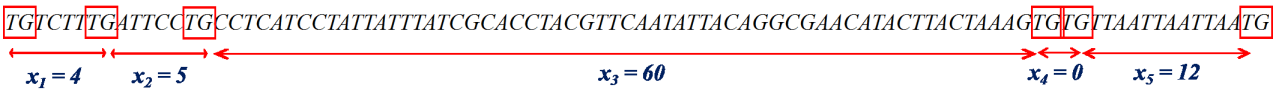

Figura 4.3: Representação esquemática para o cálculo da distância, em unidades de nucleotídeos, entre ocorrências sucessivas para o par TG de um fragmento da sequência do *Homo Sapiens*.

Tabela 4.1: Tabela das 30 primeiras sequências mitocondriais utilizadas no nosso método com seus respectivos códigos do GenBank e tamanho completo da série. No grupo 1 temos as espécies dos invertebrados e no grupo 2 temos peixes (3), anfíbios (3) e répteis (9).

Grupo 1 k	Sequência	Código GenBank	Comprimento
1	*Caenorhabditis elegans (verme1)*	*NC*_001328	13795
2	*Lumbricus terrestris (verme2)*	*NC*_001673	14999
3	*Ascaris suum (verme3)*	*NC*_001327	14283
4	*Apis mellifera (abelha1)*	*NC*_001566	16343
5	*Apis cerana (abelha2)*	*NC*_014295	15895
6	*Apis florea (abelha3)*	*NC*_021401	17694
7	*Ceratitis capitata(mosca1)*	*NC*_000857	15980
8	*Drosophila melanogaster (mosca2)*	*NC*_001709	19517
9	*Drosophila yakuba (mosca3)*	*NC*_001322	16019
10	*Pristomyrmex punctatus (formiga1)*	*NC*_015075	16180
11	*Solenopsis richteri (formiga2)*	*NC*_014677	15560
12	*Solenopsis invicta (formiga3)*	*NC*_014672	15549
13	*Strongylocentrotus purpuratus (ouriço1)*	*NC*_0014653	15650
14	*Paracentrotus lividus (ouriço2)*	*NC*_001572	15696
15	*Mesocentrotus nudus (ouriço3)*	*NC*_020771	15709
Grupo 2 k			
16	*Formosania lacustris (peixe1)*	*NC*_001727	16558
17	*Cyprinus carpio (peixe2)*	*NC*_001606	16575
18	*Rhodeus suigensis (peixe3)*	*NC*_013709	16733
19	*Xenopus laevis (sapo1)*	*NC*_001573	17553
20	*Nanorama pleskei (sapo2)*	*NC*_016119	17660
21	*Ranodon sibiricus (sapo3)*	*NC*_004021	16418
22	*Micrurus fulvius (cobra1)*	*NC*_013481	17506
23	*Naja atra (cobra2)*	*NC*_011389	17216
24	*Ramphotyphlops braminus (cobra3)*	*NC*_010196	16397
25	*Alligator mississippiensis (jacaré1)*	*NC*_001922	16646
26	*Alligator sinensis (jacaré2)*	*NC*_004448	16746
27	*Osteolaemus tetrapis (jacaré3)*	*EF*551001	16873
28	*Chelonia mydas (tartatuga1)*	*NC*_000886	16497
29	*Chelodina rugosa (tartatuga2)*	*NC*_015986	16582
30	*Pyxidea mouhotii (tartatuga3)*	*NC*_010973	16837

4.4 Resultados

Como discutido anteriormente, o conjunto W_n de todas as possíveis palavras w_n^i de tamanho n é constituído por $i = \{1, 2, ..., 4^n\}$ elementos distintos. De modo geral estas

Tabela 4.2: Tabela das 30 últimas sequências mitocondriais utilizadas no nosso método com seus respectivos códigos do GenBank e tamanho completo da série. No grupo 3 temos as espécies das aves e no grupo 4 apenas os mamíferos.

Grupo 3 k	Sequência	Código GenBank	Comprimento
31	*Meleagris gallopavo (pavão)*	NC_010195	16719
32	*Gallus gallus (galo)*	NC_001323	16775
33	*Struthio camelus (avestruz)*	NC_002785	16595
34	*Rhea americana (ema)*	NC_000846	16714
35	*Ciconia boyciana (cegonha)*	NC_002196	17622
36	*Anas platyrhynchos (pato1)*	$EU009397$	16604
37	*Dryocopus pileatus (pica-pau)*	NC_008546	16832
38	*Coturnix japonica (codorna)*	NC_003408	16697
39	*Cairina moschata (pato2)*	$EU755254$	16610
40	*Columba livia (pombo)*	NC_013978	17229
41	*Falco peregrinus (falcão)*	$JX029991$	17527
42	*Anser fabalis (ganso)*	NC_016922	16688
43	*Eudyptula minor (pinguin)*	NC_004538	17611
44	*Garrulax canorus (cuco-chines)*	NC_020429	17785
45	*Penelopides panini (bicos-de-corno)*	NC_015087	22737
Grupo 4 k			
46	*Balaenoptera physalus (baleia1)*	NC_001321	16398
47	*Globicephala macrorhynchus (baleia2)*	NC_019578	16387
48	*Cephalorhynchus heavisidii (golfinho1)*	NC_020696	16371
49	*Tursiops truncatus (golfinho2)*	NC_012059	16388
50	*Bos taurus (boi)*	$AF492351$	16338
51	*Ovis aries (ovelha)*	NC_001941	16616
52	*Capra hircus (cabra)*	NC_005044	16643
53	*Equus asinus (burro)*	NC_001788	16670
54	*Equus caballus (cavalo)*	NC_001640	16660
55	*Rattus norvegicus (rato1)*	$KF011917$	16310
56	*Mus musculus (rato2)*	NC_005089	16299
57	*Gorilla gorilla (gorila)*	NC_001645	16364
58	*Pan paniscus (bonobo)*	NC_001644	16563
59	*Pan troglodytes (chimpazé)*	NC_001643	16554
60	*Homo Sapiens (homem)*	$KC603863$	16568

sequências de letras estão presentes no DNA mitocondrial com diferentes frequências de ocorrência. Como ilustração, na Figura 4.4 apresentamos os gráficos das frequências relativas f_n^{ik} das palavras de tamanho $n = \{1, 2, 3, ..., 6\}$, para o *Homo Sapiens* (sequência $k = 60$ da Tabela 4.2). Em cada sub-gráfico (A-F) a linha em vermelho corresponde a probabilidade

de ocorrência caso as palavras fossem equiprováveis, indicando que existe uma dispersão em torno deste valor.

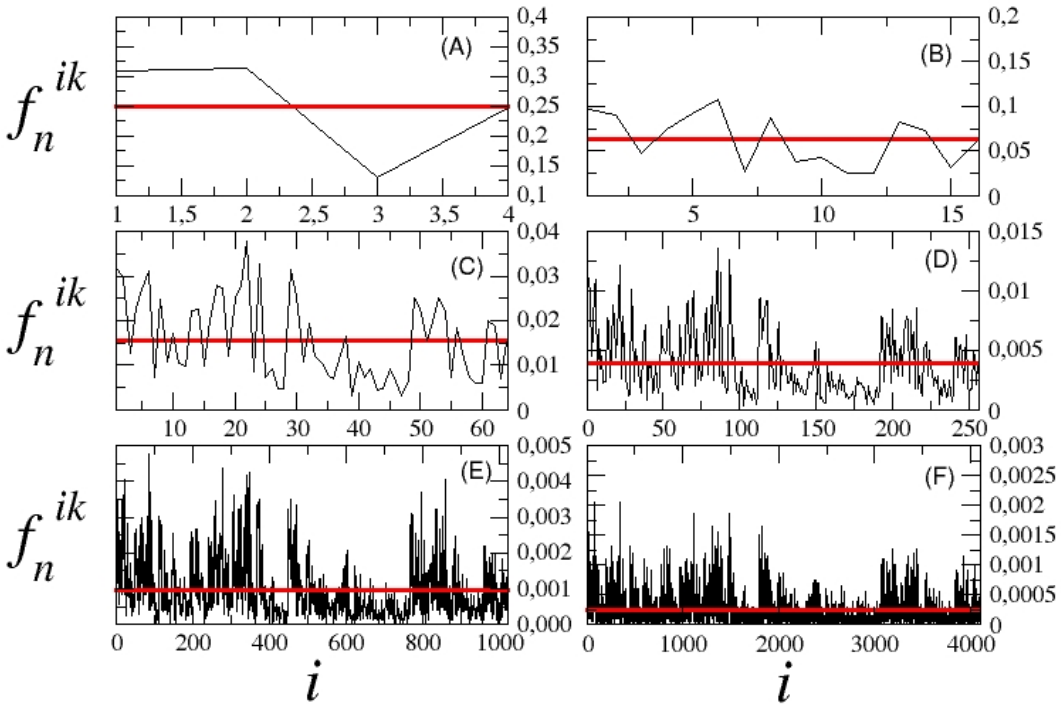

Figura 4.4: Gráfico das frequências de ocorrência f_n^{i60} das $i = \{1, 2, ..., 4^n\}$ distintas palavras presentes na sequência do DNA mitocondrial humano ($k = 60$), os sub-gráficos de (A-F) correspondem a distintos valores de tamanho $n = \{1, 2, ..., 6\}$ de palavras. Em cada caso, a linha contínua em vermelho representa a frequência caso as palavras fossem equiprováveis.

Embora possam existir 4^n palavras distintas de tamanho n, o número de elementos aos quais podemos associar o parâmetro σ, que estabelece a relevância deste elemento numa análise linguística, corresponde a uma fração deste conjunto, compreendendo palavras que ocorrem ao menos 3 vezes ao longo da sequência. A título de ilustração apresentamos na Figura 4.5 os valores de σ_n^{ik} associados a cada uma das palavras de tamanho n para o *Homo Sapiens*, assim como na Figura 4.4 nos sub-gráficos (A-F) o eixo horizontal exibe as $N \leqslant 4^n$ palavras, enquanto que no eixo vertical temos seus respectivos valores de σ. As linhas contínuas vermelhas correspondem aos valores médios dos conjuntos sobre as palavras

computáveis,

$$\overline{\sigma}_n^k = \frac{1}{N}\sum_{i=1}^{N}\sigma_n^{ik}, \tag{4.2}$$

para uma dada sequência k (neste caso $k = 60$). Uma leitura cuidadosa destes gráficos nos permite observar que os valores de σ_n^{ik} flutuam em torno da média $\overline{\sigma}_n^k$, indicando que as palavras possuem uma distribuição espacial heterogênea. Além disso, também é possível perceber que o valor médio modifica-se como função do tamanho da palavra analisada, sendo máximo para o caso em que $n = 3$, o que aparentemente indica que estas palavras seriam as mais relevantes, em analogia com linguagens naturais. De modo a sumarizar as informações discutidas até aqui apresentamos na Tabela 4.3 os valores da frequência f_n^{ik} e de σ_n^{ik}, de todas as 16 palavras de tamanho $n = 2$, para a sequência do DNA humano ($k = 60$).

Tabela 4.3: Tabela das 16 possíves palavras w_n^i de tamanho $n = 2$ para a sequência do *Homo Sapiens* ($k = 60$) com suas respectivas frequências f_n^{ik} e desvios relativos σ_n^{ik}.

i	w_2^i	f_n^{i60}	σ_n^{i60}
1	AA	$0,105$	$1,048$
2	AC	$0,098$	$0,961$
3	AG	$0,052$	$1,081$
4	AT	$0,081$	$0,948$
5	CA	$0,100$	$1,007$
6	CC	$0,116$	$0,998$
7	CG	$0,028$	$1,016$
8	CT	$0,094$	$0,949$
9	GA	$0,040$	$1,005$
10	GC	$0,047$	$1,043$
11	GG	$0,028$	$1,128$
12	GT	$0,027$	$1,217$
13	TA	$0,090$	$0,929$
14	TC	$0,079$	$1,161$
15	TG	$0,034$	$1,087$
16	TT	$0,066$	$0,993$

Reproduzindo este procedimento para todas as sequências presentes nas Tabelas 4.1 e 4.2 podemos construir um gráfico de $\overline{\sigma}_n^k$ para cada uma das espécies, como apresentado na Figura 4.6, onde o código de cores especifica os diferentes tamanhos de palavras preto ($n = 1$), vermelho ($n = 2$), verde ($n = 3$), azul ($n = 4$), amarelo ($n = 5$), marron ($n = 6$)

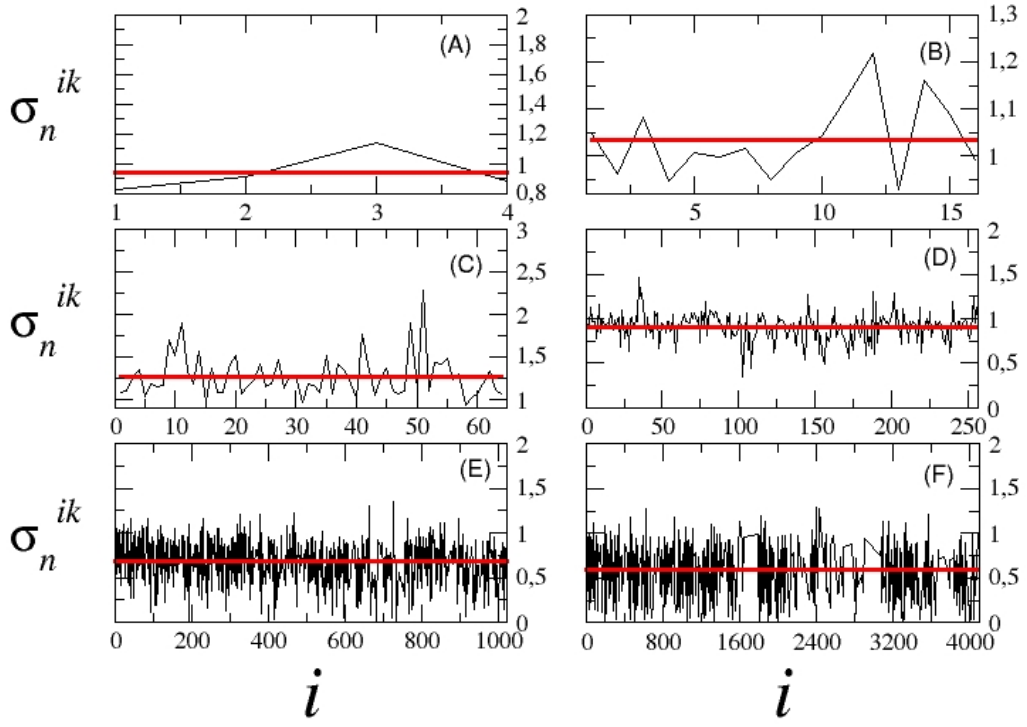

Figura 4.5: Valores de σ_n^{i60} associados a palavras de tamanho n, para a sequência do DNA mitocondrial humano ($k = 60$). Nos sub-gráficos de (A-F) variamos o valor de $n = \{1, 2, ..., 6\}$, no eixo horizontal temos as $N \leqslant 4^n$ palavras para cada caso, enquanto que no eixo vertical temos os valores de σ_n^{ik}, as linhas contínuas vermelhas correspondem aos valores médios $\overline{\sigma}_n^k$ do conjunto de palavras.

e turquesa ($n = 7$). Assim, como no caso do DNA humano, percebemos que para todas as sequências há uma tendência de que $\overline{\sigma}_n^k$ seja maximizado no caso em que $n = 3$ (curva em verde). De modo a quantificar esta característica apresentamos no sub-gráfico desta mesma figura, a média amostral

$$\langle \overline{\sigma}_n \rangle = \frac{1}{60} \sum_{k=1}^{60} \overline{\sigma}_n^k \tag{4.3}$$

sobre diferentes espécies, no qual podemos observar que de fato as palavras de tamanho

$n = 3$ apresentam os maiores valores médios de σ e corresponderiam em analogia com textos naturais à palavras mais relevantes. É interessante observar que, do ponto de vista biológico, as $N = 4^3 = 64$ palavras de tamanho $n = 3$, são identificadas como os *códons* que codificam a produção dos 20 aminoácidos que constituem todas as proteínas, como assinalado no Capítulo 2.

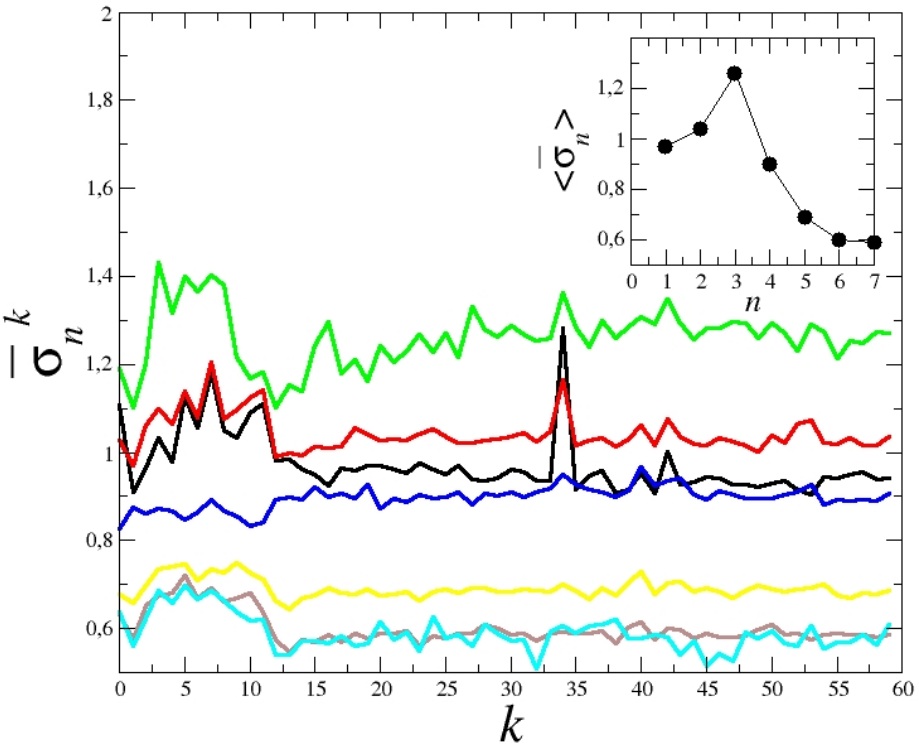

Figura 4.6: Valores de $\overline{\sigma}_n^k$ para cada uma das 60 sequências indicadas nas Tabelas 4.1 e 4.2. As cores correspondem aos diversos tamanhos n de palavras analisadas a saber: preto($n = 1$), vermelho($n = 2$), verde($n = 3$), azul ($n = 4$), amarelo ($n = 5$), marron($n = 6$) e turquesa ($n = 7$). No sub-gráfico, apresentamos a média $\langle \overline{\sigma}_n \rangle$ sobre todas as sequências como função do tamanho da palavra.

Assim, como apontado na Figura 4.4, para o caso da sequência humana, os valores de σ_n^{ik} flutuam em torno da média $\overline{\sigma}_n^k$ sobre as palavras de tamanho n de uma dada sequência k. Esta flutuação indica que a distribuição de distâncias é específica para cada palavra e que portanto elas são utilizadas em diferentes contextos. Estatisticamente esta flutuação pode ser computada por meio do desvio,

$$\Delta\overline{\sigma}_n^k = \sqrt{\overline{(\sigma_n^{ik})^2} - (\overline{\sigma_n^{ik}})^2}. \tag{4.4}$$

A Figura 4.7, apresenta o gráfico do desvio em relação a média para todas as $k = \{1, 2, ..., 60\}$ sequências e para diversos comprimentos n de palavra utilizada, com o mesmo código de cores utilizado na Figura 4.6. No sub-gráfico da Figura 4.7, exibimos a média amostral dos desvios sobre todas as sequências, como função do tamanho n das palavras, definido na equação 4.5.

$$\langle \Delta\overline{\sigma}_n \rangle = \frac{1}{60} \sum_{k=1}^{60} \Delta\overline{\sigma}_n^k \, . \tag{4.5}$$

Os gráficos apresentam uma tendência onde as palavras de maior comprimento e portanto mais raras possuem uma maior largura da distribuição dos valores de σ_n^{ik}, excetuando o caso das palavras de tamanho $n = 2$ e $n = 4$. Além disso, as primeiras 12 sequências que correspondem à espécies categorizadas dentro do grupo dos invertebrados, possuem uma dispersão maior quando comparadas às demais para o mesmo tamanho n da palavra analisada.

De forma a caracterizar o papel das palavras relevantes dentro das sequências de DNA mitocondrial, passamos à próxima fase de nosso estudo onde iremos inicialmente verificar a quantidade destas palavras em cada sequência. Na Figura 4.8, apresentamos um gráfico da fração p_n^k de palavras de comprimento n em cada sequência k que possuem $\sigma_n^{ik} > 1$, aqui as cores correspondem aos diversos tamanhos com o mesmo código das figuras anteirores. No sub-gráfico temos a média amostral sobre sequências para um mesmo valor de comprimento, indicando novamente que para $n = 3$ a fração é mais expressiva e que praticamente todas as palavras desempenhariam papel relevante no armazenamento da informação gênica.

Tendo em vista os diferentes trabalhos que se dividem entre os que defendem uma possível relação entre a linguagem natural com uma aparente "linguagem genética" [8], [7], [30], e aqueles que refutam tais proposições [10], pretendemos, então, aprofundar essa discussão oferecendo elementos que possam esclarecer ou pelo menos indicar certas limitações quando se quer agrupar diferentes espécies tratando o DNA como um texto. Vale acrescentar,

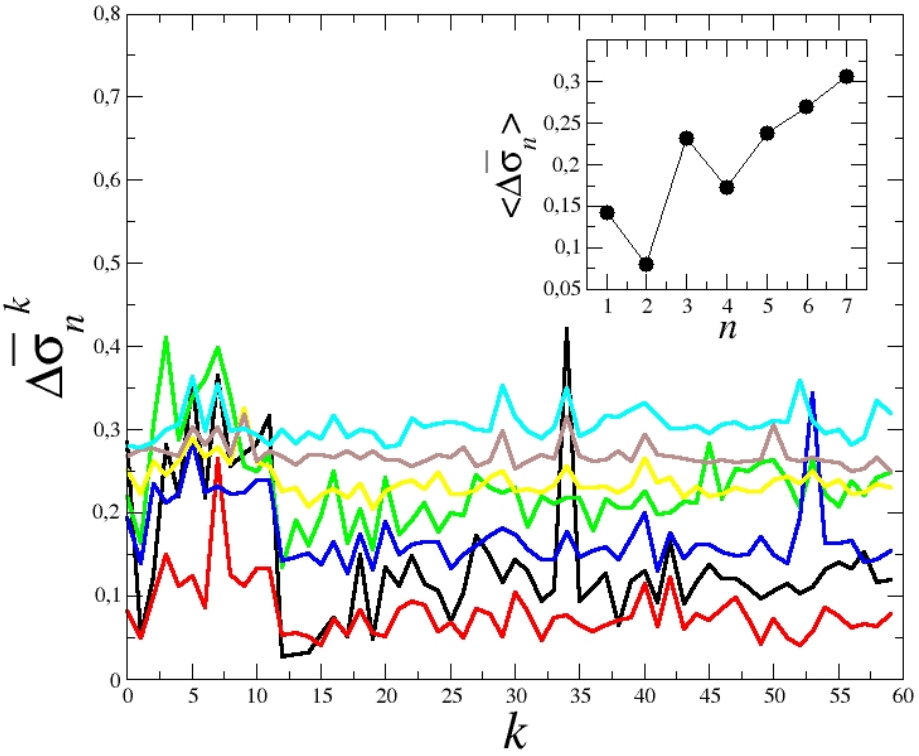

Figura 4.7: Valores de $\Delta\overline{\sigma}_n^k$ para cada uma das 60 sequências indicadas nas Tabelas 4.1 e 4.2. As cores correspondem aos diversos tamanhos n de palavras analisadas a saber: preto($n = 1$), vermelho($n = 2$), verde($n = 3$), azul ($n = 4$), amarelo($n = 5$), marron($n = 6$) e turquesa($n = 7$). No sub-gráfico apresentamos a média $\Delta\overline{\sigma}_n$ sobre todas as sequências como função do tamanho da palavra.

que existem outros trabalhos que defendem que análises detalhadas das distribuições de frequências para palavras com diferentes tamanhos, possam sugerir uma possível captura de padrões biológicos estruturais e, também, que esse possa ser um caminho provável para o estudo filogenético das espécies [35].

De forma semelhante, outros autores, apontam que o estudo da frequência de palavras pode ser conveniente para estudar sequências de um mesmo organismo ou de organismos que formam grupos biológicos [36]. Fundamentados nessas investigações, em nossos estudos consideraremos, também, análises baseadas em frequências.

Para tanto, considere por exemplo um vetor no espaço 4^n dimensional cujas entradas correspondam às frequências de cada uma das palavras de tamanho n. Vamos definir a distância

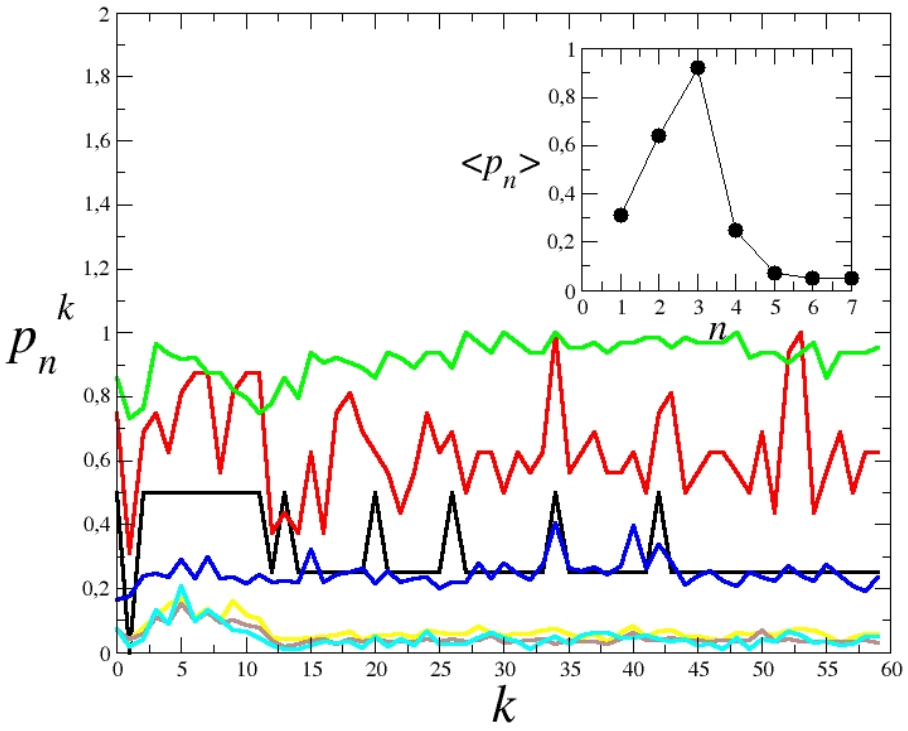

Figura 4.8: Fração de palavras relevantes p_n^k para cada uma das $k = \{1, 2, ..., 60\}$ sequências indicadas nas Tabelas 4.1 e 4.2. As cores correspondem aos diversos tamanhos de palavras n analisadas a saber: preto($n = 1$), vermelho($n = 2$), verde($n = 3$), azul ($n = 4$), amarelo($n = 5$), marron($n = 6$) e turquesa ($n = 7$). No sub-gráfico apresentamos a média $\langle p_n \rangle$ sobre todas as sequências como função do tamanho da palavra.

d_n^{lm} entre duas destas sequências l e m contidas neste espaço por meio das frequências relativas de ocorrência das palavras, como se segue:

$$d_n^{lm} = \sqrt{\sum_{i=1}^{4^n} (f_n^{il} - f_n^{im})^2} \, , \tag{4.6}$$

com esta definição podemos estabelecer uma relação de proximidade entre as sequências em termos de suas frequências. A título de ilustração, apresentamos na Tabela 4.4 as 15 sequências mais próximas daquela associada ao *Homo Sapiens* ($m = 60$), ordenadas em termos das distâncias d_n^{lm} para o caso em que $n = 2$. Como podemos observar neste caso

específico, as 3 espécies mais próximas ao homem correspondem aos outros primatas, vemos ainda que neste grupo apresentado, apenas 4 espécies encontram-se dentro da categoria dos mamíferos (grupo 4) ao qual o homem também pertence.

Tabela 4.4: Tabela das 15 sequências mais próximas do *Homo Sapiens* ($m = 60$) segundo o critério de frequências, com respectivas distâncias, levando em conta palavras de tamanho $n = 2$. Na última coluna exibimos os grupos aos quais cada espécie pertence.

l	$d_2^{l60} \times (10^{-2})$	Sequência	Grupo
59	$0,77286204$	chimpanzé	4
57	$1,04960464$	gorila	4
58	$1,10721579$	bonobo	4
31	$1,59924775$	pavão	3
38	$1,61714554$	codorna	3
43	$1,88966487$	pinguin	3
32	$1,93330925$	galo	3
23	$2,06459370$	cobra2	3
33	$2,10804306$	avestruz	3
40	$2,20620185$	pombo-comum	3
53	$2,29354799$	burro	4
25	$2,49859840$	jacaré1	2
26	$2,55801994$	jacaré2	2
22	$2,59663183$	cobra1	2
35	$2,66495850$	cegonha	3

A fim de quantificarmos qual a porcentagem de acertos na categorização podemos introduzir um parâmetro α_n^m que é a taxa de aglomeração da sequência m associada a palavras de comprimento n. Esta grandeza corresponde à fração das 15 sequências mais próximas à série m e que pertencem ao mesmo grupo desta. Seguindo este procedimento para as demais sequências contidas nas Tabelas 4.1 e 4.2 podemos definir uma taxa de aglomeração média,

$$\langle \alpha_n \rangle = \frac{1}{60} \sum_{m=1}^{60} \alpha_n^m \tag{4.7}$$

que depende apenas do comprimento n da palavra de interesse. Na Figura 4.9, apresentamos o comportamento da taxa de aglomeração média $\langle \alpha_n \rangle$, para diversos tamanhos de palavras. A primeira característica importante do gráfico é que as duas grandezas estão positivamente

correlacionadas, ou seja, existe um crescimento monotônico de $\langle \alpha_n \rangle$ com n. O segundo ponto é que palavras de tamanho $n = 3$, não exibem qualquer relevância segundo um critério "linguístico", aparentemente não desempenham qualquer papel na categorização.

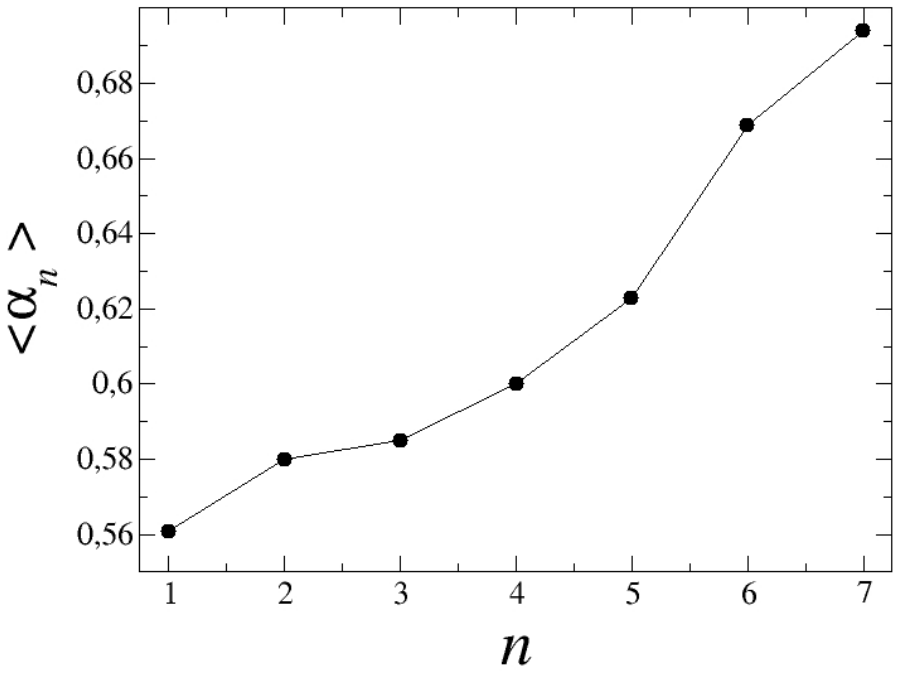

Figura 4.9: Taxa de aglomeração média $\langle \alpha_n \rangle$ como função do tamanho n da palavra considerada, observe que de forma global o parâmetro cresce monotonicamente.

Uma segunda análise, consiste em quantificar a aglomeração categorizando as espécies pelo número de palavras que possuem $\sigma_n^{ik} > 1$. Este estudo pode ser desenvolvido para cada um dos grupos das Tabelas 4.1 e 4.2 definindo uma nova taxa de aglomeração α_σ, nos mesmos moldes da anterior, diferenciando-se apenas pelo fato de que a ordenação de proximidade é dada pela interseção entre os conjuntos de todas as palavras computáveis de uma determinada sequência com as demais. Na Figura 4.10, apresentamos esta taxa de aglomeração, seguindo este procedimento. Primeiramente notamos que as taxas de aglomeração são sempre inferiores àquelas obtidas pela metodologia anterior, no qual apenas as frequências eram levadas em conta. Em seguida, percebemos que este parâmetro apresenta um valor máximo para palavras de tamanho $n = 4$ e não em $n = 3$ diferentemente do esperado.

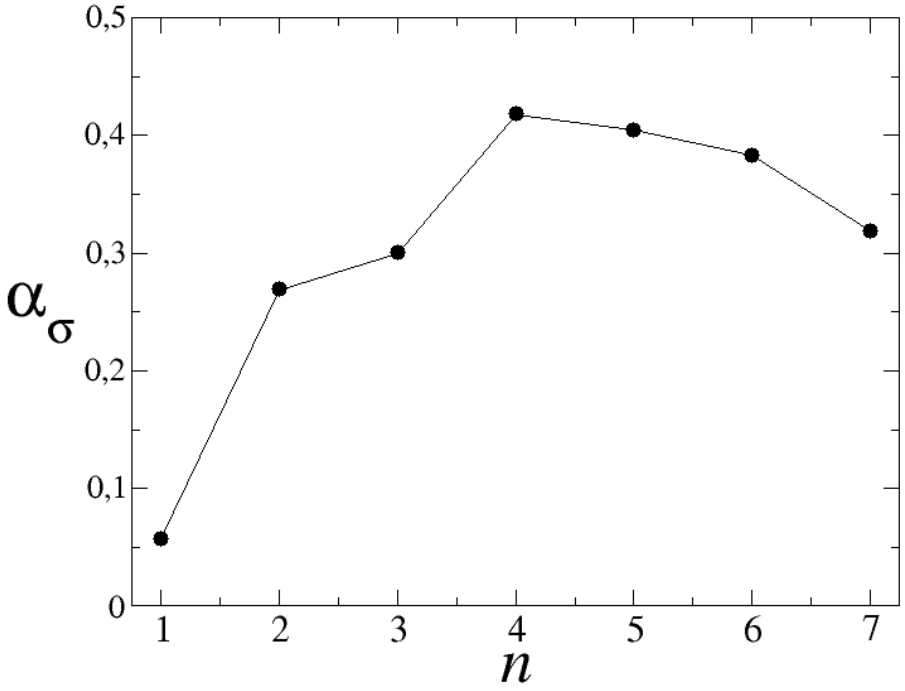

Figura 4.10: Taxa de aglomeração média α_σ como função do tamanho das palavras n cujos desvios σ são maiores do que 1, observe que temos maior aglomeração para palavras de tamanho $n = 4$.

As duas abordagens anteriores tendiam a levar em conta separadamente a frequência f_n^{ik} e a relevância σ_n^{ik} das palavras num possível método de categorização. De modo a investigar o efeito destas duas propriedades combinadas, definimos uma terceira taxa de aglomeração, $\langle \alpha_{n\sigma} \rangle$, semelhante à primeira, ou seja que leva em conta as frequências. Contudo, para o cálculo das distâncias, consideramos apenas as palavras que possuem $\sigma_n^{ik} > 1$. Na Figura 4.11, notamos que a maior taxa de aglomeração entre as espécies acontece para o caso $n = 3$. No entanto, vale destacar que, para esta abordagem conseguimos uma aglomeração em torno de $0,57$, enquanto que para a primeira, o valor da nossa taxa de aglomeação foi acima de $0,57$ no caso em que $n = 3$, o que indica que mesmo para o caso em que praticamente todas as palavras são relevantes, este cálculo da distância leva a uma menor taxa de aglomeração.

Os resultados aqui apresentados sugerem que as frequências das palavras de maior tamanho desempenham papel proeminente na categorização das sequências e portanto num possível

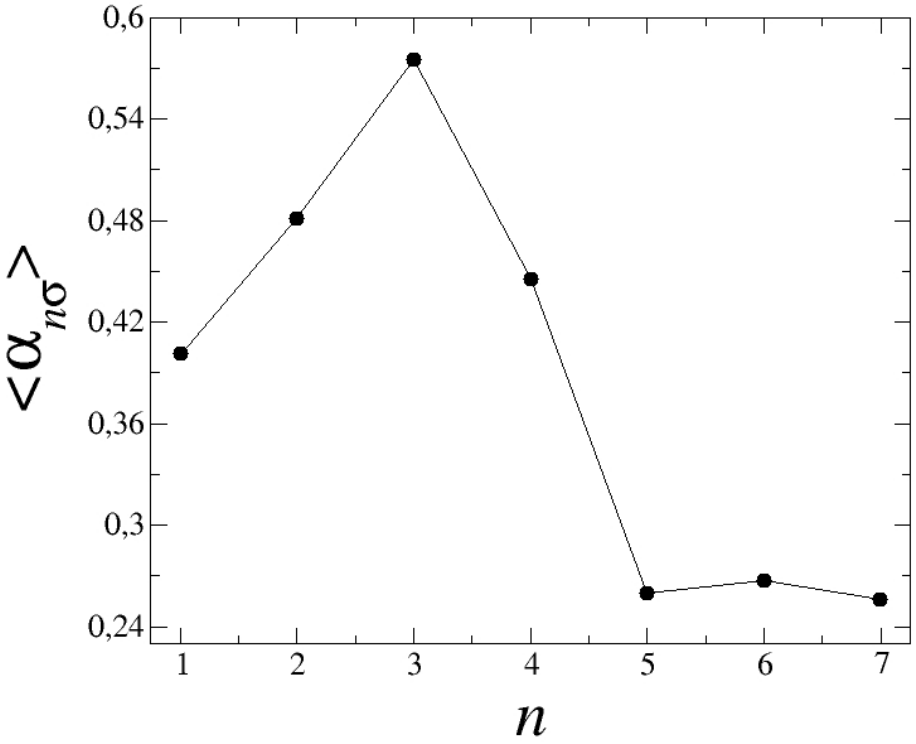

Figura 4.11: Taxa de aglomeração média $\langle \alpha_{n\sigma} \rangle$ como função do tamanho da palavra n considerada.

estabelecimento de uma árvore filogenética. Por outro lado, a distribuição espacial dessas palavras caracterizada pela variável σ_n^{ik} indica que suas posições relativas influenciam de maneira secundária. Uma vez que esta característica se manifesta mesmo para as palavras de maior tamanho analisadas, este fato dá suporte a uma possível conclusão de que o contexto em que as palavras aparecem não é relevante e que portanto um tratamento linguístico do DNA mitocondrial mereça uma análise mais cautelosa.

Capítulo 5

Conclusões e Perspectivas

Nesta dissertação, estudamos similaridades estatísticas associadas a diversas sequências completas de DNA mitocondriais. As propriedades estatísticas derivadas das distribuições de frequências de palavras ao longo das sequências permitiram capturar informações sobre conteúdos estruturais das mesmas, individualmente, bem como estimar afinidades biológicas entre diferentes espécies.

Nossos resultados confirmam a relevância do papel dos códons, como característica intrínseca das sequências codificadas e mostram que os melhores agrupamentos entre espécies distintas não ocorrem apenas em torno das palavras mais relevantes das sequências, uma vez que os melhores agrupamentos foram os obtidos calculando-se a taxa de aglomeração ou índice de similaridade levando em conta apenas as frequências das palavras. Observamos ainda que quanto maior o tamanho da palavra mais eficaz é o agrupamento entre as sequências.

A metodologia empregada em nossos estudos se constitui em uma alternativa às técnicas e algoritmos de alinhamento de sequências, cujos resultados são satisfatórios apenas para sequências pequenas, pois são de difícil aplicação para sequências grandes e de tamanhos variáveis.

Além disso, a utilização de métodos baseados em frequências reduz substancialmente

a dimensão das sequências, retendo somente a informação estrutural que possui relevência biológica, como, por exemplo, o papel dos códons. Evidentemente que outras aplicações se fazem necessárias em grupos de sequências da mesma espécie (por exemplo, bactérias) para confirmarmos o cenário acima e inferir outras informações biológicas.

No mais, como perspectivas, poderemos realizar novos testes para investigar se os resultados obtidos aqui, no contexto de sequências genômicas, apresentam correlação com os aspectos estatísticos típicos de textos linguísticos e sequências simbólicas, pois a hipótese sobre a existência de uma linguagem genética permanece um problema em aberto. Finalmente, a possibilidade de aplicação de nossos resultados pode ser relevante na construção de árvores filogenéticas que são estruturas adequadas para se compreender a história evolucionária dos organismos.

Referências Bibliográficas

[1] GenBank Overview. http://www.ncbi.nlm.nih.gov/genbank/. Acessado em agosto 2013.

[2] The Worldwide Protein Data Bank. http://www.wwpdb.org/. Acessado em agosto 2013.

[3] E.Schrodinger. O que é vida? - O Aspecto Físico da célula viva. São Paulo: Editora UNESP, 1ª edição, 1997.

[4] J. D. Watson. DNA - O Segredo da Vida. São Paulo: Editora Companhia das Letras, 1ª edição, 2005.

[5] J. F. Miescher. http://www.scienceinschool.org/2006/issue1/discoveringdna. Acessado em agosto 2013.

[6] J. D. Watson e F. Crick. Molecular Structure of Nucleic Acids. Nature, 171:737–738, 1953.

[7] R. N. Mantegna; A. L. Goldberg; S. V. Buldurev; S. Havlin; C. -K. Peng; M. Simons; H. E. Stanley. Linguistic Features of Noncoding DNA Sequences. Phys. Rev. E, 73:3169–3172, 1994.

[8] M. Ortuño; P. Carpena; P. Bernaola-Galván; E. Muñoz; A. M. Somoza. Keyword Detection in Natural Languages and DNA. Europhys. Lett., 57, 759–764, 2002.

[9] J. A. Glazier; S. Rachavachari; C. L. Berthelses; M. H. Skolnick. Reconstructing Phylogeny from the Multifractal Spectrum of Mitochondrial DNA. Phys. Rev. E, 51, 2665–2668, 1995.

[10] A. A. Tsonis, J. B. Elsner e P. A. Tsonis. Is DNA a Language?. J. Theor. Biol., 184, 25–29, 1996.

[11] Nucleotideo. http://scienceblogs.com/. Acessado em agosto 2013.

[12] Dupla Helice. http://scienceblogs.com/. Acessado em agosto 2013.

[13] Discovery of DNA Structure and Function: Watson and Crick. http://www.nature.com/scitable. Acessado em agosto 2013.

[14] Tabela dos Aminoácidos. http://www.nature.com/scitable. Acessado em agosto 2013.

[15] D. C. Wallece. Diseases of the mitochondrial DNA. Annu. Rev. Biochem, 61, 1175–1212, 1992.

[16] S. Anderson and A. T. Bankier and B. G. Barell and M. H. L. Bruijn and A. R.Coulson and J. Drouin and I. C. Eperon. Sequence and organization of the human mitochondrial genome. Nature, 290, 457–465, 1981.

[17] G. C. Kujoth; A. Hiona; T. D. Pugh; S. Someya; K. Panzer; S. E. Wohlgemuth; T. Hofer; A. Y. Seo; R. Sullivan; W. A. Jobling; J. D. Morrow. Mitochondrial DNA Mutations, Oxidative Stress, and Apoptosis in Mamalian Aging. Science, 309, 481–484, 2005.

[18] S. DasSarma. Extreme Halophiles are Models for Astrobiology. Microbe, 1, 120–126, 2006.

[19] P. M. Gleiser, S. A. Cannas e F. A. Tamarit. Long-Range Effects in Granular Avalanching. Phys. Rev. E, 63, 042301, 2001.

[20] M. V. Carneiro e I. C. Charret. Criticalidade Auto-Organizada na Pilha de Areia. Rev. Bras. Ens. Fis., 27, 571–576, 2005.

[21] R. J. Geller. Earthquake Prediction: A Critical Review. Geophys. J. Int., 131, 452–450 1997.

[22] R. V. Solé. On Macroevolution, Extinctions and Critical. Complexity, 1, 40–44, 1996.

[23] C. -K. Peng; S. V. Buldyrev; A. L. Goldberger; S. Havlin; F. S. Ciortino; M. Simons; H. E. Stanley. Long-Range Correlations in Nucleotide Sequences. Nature, 356, 168–170, 1992.

[24] C. -K. Peng; S. V. Buldyrev; S. Havlin; M. Simons; H. E. Stanley; A. L. Goldberger. Mosaic Organization of DNA Nucleotides. Phys. Rev. E, 49, 1685–1689, 1994.

[25] A. Rosas, E. Nogueira Jr. e J. F. Fontanari. Multifractal Analysis ad DNA Walks and Trails. Phys. Rev. E, 66, 061906, 2002.

[26] X. Gu e W. -H. Li. A model for the correlation of mutation rate with GC content and the origin GC-rich isochores. J. Mol. Evol., 38, 468–475, 1994.

[27] W. -H. Li. Expansion-Modification Systems: A Model for Spatial 1/f Spectra. Phys. Rev. A, 43, 5240–5260, 1991.

[28] H.E. Stanley; V. Afanasyev; L.A.N. Amaral; S.V. Buldyrev; A.L. Goldberger; S. Havlin; H. Leschhorn; p. Maass; R.N. Mantegna; P.A. Prince; M.A. Salinger. Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. Phys. Rev. A, 224, 302–321, 1996.

[29] G. K. Zipf. Human Behavior and the Principle of Least Effort. Cambridge, 1949.

[30] R. N. Mantegna; S. V. Buldyrev; A. L. Goldberger; S. Havlin; C. -K. Peng; M. Simons; H. E. Stanley. Systematic analysis of coding and noncoding DNA sequences using methods os statistical linguistics. Phys. Rev. E, 52, 2939–2950, 1995.

[31] C. E. Shannon. A mathematical theory of communication. Bell Syst. Tech. J., 27, 1948.

[32] T. A. Brody; J. Flores; J. B. French; P. A. Mello; A. Pandey; S. S. M. Wong. Random-matrix physics: spectrum and strength fluctuations. Rev. Mod. Phys., 53, 385–480, 1981.

[33] N. N. Oiwa e J. A. Glazier. The fractal structure of the mitochondrial genomes. Physica A., 311, 221–230, 2002.

[34] M. Hackenberg; A. Rueda; P. Carpena; P. Bernaola-Galván; G. Barturen; J. L. Oliver. Clustering of DNA Words and Biological Function: A Proof of Principle. Jour. Theor. Biol., 297, 127–136, 2012.

[35] P. Chaudhuri e S. Das. Statistical analysis of large DNA sequences using distribution of DNA words. Current Science, 80, 1161–1166, 2001.

[36] S. Basu, D. P. Burma e P. Chaudhuri. Words in DNA Sequences: Some Case Studies Based on their Frequency Statistics. J. Math. Biol., 16, 479–503, 2003.