Estimadores de Entropia para Seqüências de DNA

José Osvaldo Couto Horta

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO

PARA OBTENÇÃO DO GRAU DE MESTRE
EM
MATEMÁTICA APLICADA

Área de Concentração : Ciência da Computação
Orientador : Prof. Dr. José Coelho de Pina Júnior

- São Paulo, junho de 2001 -

Estimadores de Entropia para Seqüências de DNA

Redação final da dissertação de mestrado,
com a implementação das correções propostas
pela comissão julgadora.

São Paulo, 31 de outubro de 2001.

Banca examinadora

Membros Titulares:

- Prof. Dr. José Coelho de Pina Júnior (Presidente) / IME - USP
- Prof. Dr. José Augusto Ramos Soares / IME - USP
- Prof. Dr. João Meidanis / IC - UNICAMP

Membros Suplentes:

- Profa. Dra. Nami Kobayashi / IME - USP
- Prof. Dr. João Carlos Setubal / IC - UNICAMP

Agradecimentos

Ao meu pai José Célio
pelo estímulo e preocupação.
À minha mãe Maria Clélia
por sua dedicação e
empenho na revisão da escrita.
À minha esposa Cida
por sua paciência e perseverança.
À minha filha Elisa
por sua motivação e compreensão.
Aos meus irmãos Eduardo, Sérgio e Ana Clélia
pela força e confiança.
Ao orientador e amigo Coelho
que soube cobrar quando necessário e ser paciente quando eu precisei.
À professora Cristina Gomes Ferreira
por seu apoio.
Ao professor José Augusto Ramos Soares
por sua visão esclarecedora.
Ao professor Paulo Feofiloff
que teve a capacidade de mostrar ao engenheiro os caminhos da matemática.
A todas as pessoas que me auxiliaram e incentivaram no decorrer deste trabalho.

Resumo

Nesta monografia são descritas aplicações da teoria da informação à biologia molecular, mais precisamente, são apresentadas aplicações de estimadores de entropia nas seqüências de DNA. Dentre estas estão: auxílio tanto na comparação genômica quanto na construção de árvores filogenéticas e a localização de regiões do DNA com algum significado biológico específico, por exemplo, na determinação dos limites entre os íntrons e os exons. É feita uma comparação entre os diversos estimadores e uma análise da previsibilidade à distância dos genes.

Abstract

This monograph describes applications of information theory to molecular biology, more precisely, the presentation of applications of entropy estimators to DNA sequences. Among these are: genome comparison, the construction of phylogenetic trees and the detection of DNA regions with a specific biological meaning as the intron/exon boundaries. A comparison of these estimators is presented and so an analysis of gene at a distance predictability.

Índice

Capítulo 1

Introdução

> "Uma curiosa falta de comunicação entre especialistas em biologia molecular, matemáticos e engenheiros elétricos que desenvolveram a teoria da informação e a teoria da codificação resultou na impossibilidade dos especialistas em biologia molecular enxergarem que a solução dos seus problemas, em alguns casos, já havia sido desenvolvida. Da mesma forma os especialistas da área de exatas não conseguem ver as oportunidades muito relevantes de aplicar a teoria da informação à biologia molecular."
>
> H.P. Yockey [56]

Há diversas especulações a respeito da origem da vida [12, 35, 20, 7]. Os mecanismos de auto-replicação são requisitos essenciais nestas possibilidades. As enzimas, que são um tipo de proteína, não poderiam ser utilizadas nos mecanismos primordiais de auto-replicação, pois elas são produzidas a partir dos códigos genéticos e somente foram criadas após a vida. Um dos atuais campos de pesquisa da bioquímica está na busca de sistemas auto-replicantes que não utilizem enzimas. Caso este sistema vier a ser descoberto, haverá um ponto para os defensores da hipótese de que a vida surgiu espontaneamente a partir da sopa primordial[1] de substâncias orgânicas do oceano pré-biótico. Mesmo os defensores desta hipótese não acreditam que existiria grande quantidade de RNA na sopa primordial, pois a citosina, base nitrogenada existente tanto no DNA quanto no RNA, teria pouca possibilidade de ser produzida nas condições do oceano pré-biótico. Outra teoria supõe que a vida tenha vindo à Terra do espaço[2], colocação esta que não

[1] Stanley Miller em 1953 conseguiu produzir aminoácidos e outras substâncias orgânicas utilizados pelas células vivas a partir de descargas elétricas, em uma simulação da atmosfera primitiva do planeta Terra. Com isto se mostrou a possibilidade da vida ter origem na sopa primordial [46].

[2] Apresentada na década de 70 por Hoyle e Wickramasinghe [4, 56].

explica sua origem, nem como teria sido ativado o código genético da 'vida'[3], mas possibilitaria a utilização de enzimas nos processos de auto-replicação [57]. Para Hoyle, supor que a vida se originou na Terra não passa de uma faceta do geocentrismo, desbancado por Copérnico no século XVI [1]. Qualquer que seja sua origem, a existência de água é essencial à vida. Por mais primitiva ou evoluída que a vida se apresente, esta somente existe na presença de água no estado líquido. No capítulo 2 serão citados aspectos da importância fundamental da água para a vida.

Há outras hipóteses para a origem da vida [12, 7] e diversas controvérsias envolvendo o DNA. Mas em dois pontos o DNA é incontestável. O primeiro está no processo de auto-replicação da vida. Todos os seres vivos possuem o código genético (DNA ou RNA) e é isto que os distingue da matéria morta. Há diversos tipos de seres vivos: animais, vegetais, vírus, etc e todos invariavelmente dependem do código genético para que possam se reproduzir. Em alguns tipos de vírus o DNA não está presente; eles possuem só o RNA, mas necessitam de uma célula (que possua DNA) para sua reprodução.

Outra consideração sobre os seres vivos está na diversidade dos ambientes nos quais eles vivem, chamados *ecossistemas*. Cada ecossistema possui suas cadeias (ou teias) alimentares. Os seres autotróficos[4] são a base das cadeias alimentares. A partir destes, a energia é passada para os demais seres vivos do ecossistema [43]. O *Methanococcus jannaschii*[5] é um exemplo de organismo autotrófico. Na seção 1.3 serão descritos aspectos do ecossistema no qual ele vive. A teoria da evolução, defendida por Charles Darwin (1809-1882), propõe que os seres vivos evoluem; com o decorrer do tempo esta evolução pode produzir novas espécies. Este mecanismo (evolução) somente é possível porque existe a seleção natural. As espécies geram mais indivíduos do que os que são capazes de sobreviver. Existem pequenas variações entre os indivíduos da mesma espécie (pernas maiores, pelagem mais espessa, por exemplo). Embora a sorte possa determinar sobre a vida dos indivíduos (da mesma espécie) estas pequenas diferenças geram vantagens (ou desvantagens) que por sua vez acabam definindo quais indivíduos sobreviverão. Um predador com pernas maiores pode conseguir mais caça; outro com pelagem mais espessa pode resistir ao inverno. Somente os sobreviventes passarão aos descendentes seus genes com suas características que serão transmitidas. Se repetido por sucessivas gerações, este processo provoca a adaptação da espécie ao meio no qual ela vive. Este fato pode originar uma nova espécie, se houver tempo (e adaptações) suficiente. Um exemplo clássico de seleção natural ocorreu na Inglaterra no século XIX envolvendo a cor do tronco das árvores e um tipo de mariposas [20]. O planeta Terra está em constante mutação (a glaciação, por exemplo). Neste processo lento, as espécies vão se

[3] A entrada de um meteoro na atmosfera geraria muito calor...

[4] 'Nutrição autotrófica (auto=próprio, trofo=alimento) realizada apenas pelas plantas, algas e por certas bactérias - o organismo é capaz de produzir todas as moléculas orgânicas de seu corpo a partir de substâncias inorgânicas que retira do ambiente ...' [22].

[5] Teve seu DNA completamente seqüenciado. Seu genoma foi utilizado por Nevill-Manning e Witten [39] analisado mais à frente.

adaptando e evoluindo (alterações genéticas de longo prazo) em função do ambiente em que elas vivem.

O segundo ponto a respeito do DNA que não gera polêmica é o fato do DNA possuir informação[6]. Embora o significado biológico do seqüenciamento genético esteja longe de ser compreendido, pelo menos um pequeno pedaço do DNA (nos seres eucarióticos) tem seu significado prático conhecido: são os genes. Os genes servem para codificação das proteínas. Enquanto na bactéria *Escherichia coli* quase todo seu código genético é utilizado na codificação, na planta *fritillaria* apenas 0,02% o é [39]. Além disto, principalmente em genomas virais, há casos de *superposição de genes*, um tipo de organização genética que permite que uma mesma seqüência codifique duas proteínas diferentes entre si; a superposição pode ser parcial ou total [58]. Embora seja de uma forma não inteiramente conhecida e variando de genoma para genoma, é certo que o DNA guarda informações sobre as proteínas sintetizadas pelas células que o contêm.

A informação depende de uma mensagem e esta depende de um transmissor, um receptor e um meio de transmissão. Segundo Shannon, o problema fundamental da comunicação é reproduzir a mensagem em algum local distinto daquele na qual ela foi produzida. As mensagens, seqüências de símbolos, são enviadas por um sistema de comunicação e possuem algum significado. As cadeias polinucleotídicas, como as mensagens, são transmitidas e também carregam acepção. Tanto os sistemas de comunicação quanto a teoria da informação tratam do envio de mensagens de um ponto a outro, do passado para o presente, do presente para o futuro. A evolução pode ser considerada como um sistema de comunicação do passado para o presente [56].

As transmissões podem sofrer influências externas as quais provocam um erro chamado de *interferência*. Embora a estrutura do DNA seja complexa, a informação hereditária nele contida é linear, como se fosse uma seqüência de caracteres onde cada um deles pode ser uma base nitrogenada, estrutura primária do DNA, ou um aminoácido, estrutura primária da proteína. A transmissão da informação do DNA para a célula é feita através do RNA mensageiro e qualquer interferência neste processo pode gerar uma proteína que não cumpra corretamente seu papel dentro da célula. Da transcrição do DNA até a síntese protéica há dois mecanismos de conferência com probabilidade de erro de 3×10^{-4}. Isto gera uma probabilidade de erro de 9×10^{-8} no processo [56]. Este número é bem menor que o inverso do tamanho de um gene (entre 1.000 e 100.000 nucleotídeos), fato que garante o bom funcionamento das células.

Quando Shannon desenvolveu seu trabalho Uma Teoria Matemática da Comunicação [49] em 1948, ele propôs o seguinte esquema genérico de sistema de comunicação, como visto na figura 1.1:

---

[6]'... Informação: medida da redução da incerteza, sobre um determinado estado de coisas, por intermédio de uma mensagem; neste sentido, informação não deve ser confundida com significado e apresenta-se como função direta do grau de originalidade, ou de valor-surpresa da mensagem, sendo quantificada em bits de informação. ...' [14].

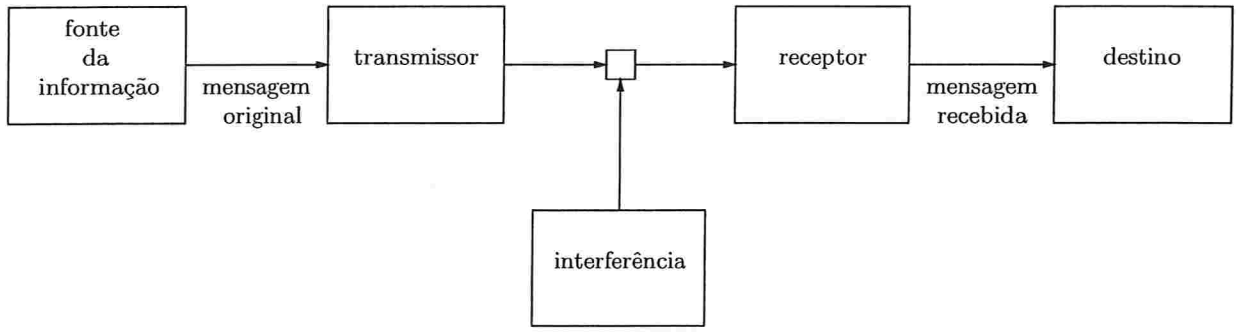


Figura 1.1: Sistema de comunicação de Shannon.

Na figura 1.2 é mostrada a transmissão da mensagem dentro da célula no processo de síntese protéica. Neste caso, a fonte de informação é o próprio DNA. A mensagem original é extraída dele através do RNA polimerase, o transmissor é o RNA mensageiro, o meio no qual pode haver a interferência é o interior da célula, o receptor é o ribossomo (que efetivamente sintetiza a proteína) e o destino é a própria célula que irá utilizar a proteína. A interferência pode ocorrer em qualquer ponto do processo de síntese protéica. Para fins de simplificação foi mantida a interferência no meio do processo.

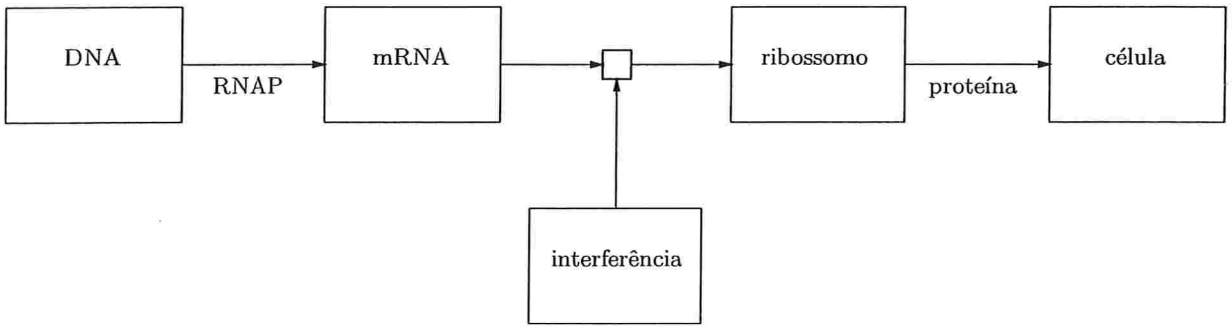


Figura 1.2: Sistema de comunicação DNA-mRNA-proteína.

No caso específico do DNA, para os exons existe o conhecimento biológico do significado da informação. Em geral este significado não é relevante para a teoria da informação, pois não há meios matemáticos de tratá-lo. O problema fundamental da comunicação está em reproduzir a mensagem original em um ponto de destino. Os mecanismos internos de transmissão de mensagens da célula garantem que este processo é executado com alto grau de confiabilidade. A entropia foi definida por Shannon como uma função de uma distribuição de probabilidade. Para definir entropia Shannon supôs que a mensagem foi modelada através de um processo de Markov.[7] Aparentemente as seqüências das proteínas não são codificadas através deste proces-

[7]Processo pelo qual há uma dependência entre um certo número de antecessores imediatos com o próximo

so [39]. Desta forma, as tentativas de compressão das proteínas falharam para os algoritmos tradicionais [31], conseguindo apenas uma pequena melhora (em torno de 5%) no tamanho da mensagem codificada. Em alguns casos o tamanho ficou maior que o previsto pela distribuição aleatória, o que de certa forma nos indica que a informação genética está bem 'comprimida' pela natureza.

Além da possibilidade de erro na codificação da proteína, o DNA está sujeito a uma segunda forma de influência: a mutação genética. Novamente há pelo menos dois mecanismos no DNA para evitar este tipo de alteração. O primeiro está em sua própria estrutura física, pois as bases nitrogenadas só combinam entre si numa ordem rígida. Como o DNA é composto por uma dupla cadeia polinucleotídica, uma mutação genética somente será possível se houver alteração nas duas cadeias. O segundo mecanismo está na redundância do código genético. Há vários aminoácidos com mais de uma forma de representação, de forma que uma mutação genética pode transformar um aminoácido em outra forma de sua própria representação, o que não alteraria a função da proteína a ser codificada. A alteração genética pode ocorrer de duas formas distintas: a alteração de uma base nitrogenada ou a inclusão (ou exclusão) de um nucleotídeo no código genético. Alguns fatos podem ocorrer quando da mutação genética, tais como: modificação de um aminoácido para algum dos códons de fim de codificação, a alteração do códon de fim de codificação em algum aminoácido, modificação de muitos aminoácidos da proteína, modificação de apenas um aminoácido ou nada. Caso ocorra a mutação genética, a alteração ocorreu na mensagem e não no transporte da informação dentro da célula. A mutação genética pode ser viável (quando permite a funcionalidade da célula) ou não. No primeiro caso ela é passada para os descendentes e, na eventualidade de representar um ganho, pode vir a ser no futuro uma forma de evolução da espécie. No último caso naturalmente não há a transmissão da mutação para os descendentes.

1.1 Importância da utilização dos estimadores de entropia no DNA

O DNA contém informação. A entropia é uma medida natural para os seguintes fenômenos: complexidade, compressibilidade, previsibilidade e aleatoriedade [19]. Este fato por si já seria suficiente para justificar a utilização de estimadores de entropia no estudo da seqüência do DNA. Porém a utilização dos estimadores de entropia transcende aos fenômenos que ela mede. Um dos principais motivos para a estimativa da entropia ser utilizada no DNA está na falta de conhecimento a respeito do próprio DNA. Exceto pelos genes (mais precisamente as regiões dos genes utilizadas na codificação, também chamadas de exons), não se sabe para que serve

símbolo da cadeia.

o DNA. Uma das mais recentes controvérsias a respeito do DNA está na quantidade dos genes humanos, pouco maior que a quantidade de genes de um rato, em torno de 30.000 no total. Em uma entrevista recente [30], a geneticista da USP Mayana Zats declarou: ' ...E teve também a "decepção" provocada pelo fato de termos "só" 30.000 genes, como se fosse pouco. ... ' Uma pequena reportagem assinada por Flávio Dieguez [15] ' ...O segundo engano refere-se ao fato de que os genes constituem apenas 3% do genoma: o resto parece inútil, tanto que está sendo chamado de "lixo genético". Mas esse material, como sugere um estudo recém-publicado pelos cientistas Carla Goldman e Nestor Oiwa, da Universidade de São Paulo, pode ter a função de organizar os genes e ordenar suas combinações. "O lixo pode ser o próprio responsável pela complexidade do organismo humano" diz Carla.' Quando se faz a estimativa da entropia de uma mensagem, normalmente é conhecida a estrutura por trás desta mensagem. Isto permite que seja criado um modelo que, retratando esta estrutura, mostre um valor realista da estimativa da entropia desta mensagem (para o modelo adotado). Mas que modelo adotar para as cadeias de DNA, se não se conhece a estrutura por trás do código genético?

Uma possível resposta a esta pergunta está nos conhecimentos prévios da teoria da informação e na compressão de dados. Como não se conhece a estrutura do DNA, são feitas várias tentativas de fazer a estimativa da entropia do DNA, através da utilização dos modelos freqüentes para a compressão de textos e dados. No capítulo 6 foi criado um programa que emprega uma variação dos algoritmos LZ (técnica de dicionário). Loewenstern e Yianilos [31] utilizaram um modelo de estados finitos com comparações inexatas visto no capítulo 7. Nevill-Manning e Witten [39] abordaram no capítulo 8 a versatilidade das técnicas de *blending* e Lanctot, Li e Yang [28] produziram um algoritmo estudado no capítulo 9 que nos remete aos modelos gramáticos. Em cada um destes artigos há uma inovação no modelo utilizado, levando em conta características dos genes e do DNA ou criando otimizações de procedimentos universais. Criar procedimentos de compressão universais é um dos menores motivos para o estudo da estimativa da entropia do DNA, embora seja ainda um desafio em aberto o aprimoramento das técnicas de modelagem e compressão de dados para seqüências de DNA e em especial nas cadeias de exons.

O principal motivo de aplicar estimadores de entropia nas cadeias de DNA está na busca de "padrões de comportamento" em sua estrutura primária. Através destes padrões de comportamento é possível buscar regiões do DNA que possuam significado próprio, como o limite íntron/exon pesquisado em Farach, et al. [19] ou mesmo buscar algum padrão e, a partir do padrão, verificar a existência ou não de significado biológico do mesmo. Outra utilização dos padrões de comportamento está na busca de correlações entre DNAs distintos, como foi feito por Chen et al. [10]. Os genomas do homem e do chimpanzé são muito similares. Ao se treinar o estimador de entropia no homem e utilizá-lo no chimpanzé ou vice-versa, as estimativas de entropia serão similares. Porém, se for treinado em mamíferos e utilizado em fermentos [31], as estimativas de entropia serão distintas. Para estas comparações parece óbvio que as semelhanças

e/ou diferenças serão marcantes. Mas suponha que se queira comparar um comportamento de um vírus com um outro para o qual já exista uma vacina. Este tipo de comparação não possui um resultado pré-determinado e este pode interferir no tipo de atitude a ser tomada na busca de uma cura para o primeiro vírus. Outro uso para os padrões de comportamento está na possível compreensão do código genético. Um efeito colateral do estudo da estimativa da entropia do DNA pode ser a descoberta de um padrão no código genético que seja geral (não necessariamente universal) e a partir daí haja uma contribuição para que os biólogos possam entender qual a serventia do lixo genético (ou um pequeno pedaço deste).

Ainda há muita incerteza envolvendo as possibilidades de utilização dos estimadores de entropia nas seqüências do DNA, mas a entropia é uma ferramenta que está sendo amplamente utilizada e tem potencial para gerar bons resultados.

1.2 Como este trabalho está estruturado

Este trabalho está dividido em quatro blocos. O primeiro destes trata da estrutura básica para a compreensão da estimativa da entropia da seqüência do DNA. O capítulo 2 mostra os aspectos da biologia molecular necessários à compreensão do DNA, RNA e proteínas. O estudo parte da célula até chegar ao DNA e às proteínas, situando desta forma o DNA no contexto ao qual ele pertence e atua. O capítulo 3 aborda a teoria da informação e suas bases probabilísticas. O enfoque parte da probabilidade, base da teoria da informação; passa pelos conceitos de compressão de dados, utilização prática da teoria da informação; termina com a entropia, parte da teoria da informação, descrevendo sem detalhar os aspectos relativos aos modelos que são vistos no capítulo 4. Os modelos foram separados em um capítulo por serem muito relevantes, tanto para a teoria da informação quanto para as técnicas de compressão de textos e dados. Finalmente são vistas as principais técnicas de compressão no capítulo 5. Embora haja um grande número de compressores, há um enfoque específico para os que foram utilizados nos artigos científicos estudados. Algum compressor que não seja citado nos artigos é mostrado por seu aspecto histórico ou pela aplicabilidade de algum modelo.

O segundo relata os artigos científicos estudados para a análise da estimativa da entropia do DNA. O artigo de Farach, et al. [19] cria um compressor baseado nos algoritmos LZ. Já o trabalho de Loewenstern e Yianilos [31] faz considerações a respeito da estimativa da entropia do DNA, além de criar um algoritmo que proporcionalmente reduziu bastante a estimativa da entropia do DNA. O artigo de Nevill-Manning e Witten [39] discorre sobre a dificuldade de tratar a estimativa da entropia da estrutura primária das proteínas, levando em consideração aspectos relativos às mutações genéticas. O trabalho de Lanctot, et al. [28] consegue melhorar a estimativa da entropia do DNA, utilizando transformações gramáticas em sua seqüência. A comparação entre genomas é feita no capítulo 10 por Chen, Kwong e Li [10].

O terceiro mostra as conclusões. No capítulo 11 são vistas as aplicações biológicas dos programas que utilizam os conceitos da teoria da informação. Há um destaque para o limite íntron/exon e para a construção das árvores filogenéticas. O capítulo 12 relata o desempenho dos compressores criados nos diversos artigos, mostrando seus pontos fortes e fracos. Os resultados da compressão são padronizados (em bits por símbolo) e mostrados em três tabelas. As experiências com previsibilidade à distância são relatadas no capítulo 13. Finalmente as conclusões finais são apresentadas no capítulo 14 onde ocorre o fechamento deste trabalho.

O último auxilia na leitura da dissertação, incluindo esta introdução, a referência bibliográfica, o glossário e o índice remissivo.

Houve um esforço para a consistência da simbologia utilizada. Eventualmente algum capítulo pode apresentar uma simbologia própria. Neste caso houve a preocupação de fazer uma definição específica cuja aplicação ficou restrita ao capítulo. Como a literatura adotada foi bastante heterogênea, em alguns casos foi necessária a representação do contexto da referência bibliográfica. As notações individualizadas servem como instrumento de clareza do texto, retratando a realidade para a qual foi escrita. Porém a notação está homogeneizada onde não houve riscos à compreensão.

1.3 Ecossistema dos ventos hidrotermais

Vivemos na superfície do planeta Terra, a uma temperatura média de 15^{o}C e com uma pressão atmosférica em torno de 760mm de mercúrio (algo equivalente a 10 metros de coluna d'água ou uma atmosfera). O sustento do ecossistema que nos alimenta é o sol. A partir de sua luz e calor, as plantas conseguem fazer a fotossíntese e se tornam a base da cadeia alimentar. A vida marinha também é sustentada pelo sol, pois seus vários seres, que realizam fotossíntese, são a base de suas diversas cadeias alimentares. Porém a luminosidade do sol tem um limite de penetração[8]. A partir desta profundidade, os diversos hábitats passam a depender de outras fontes de alimento que não sejam a luz. As regiões abissais são um exemplo deste tipo de ecossistemas, com profundidades acima dos 4.000m. Elas dependem basicamente de restos de animais mortos e fezes como base da teia alimentar [23], e estão sob uma temperatura de 2^{o}C o ano todo.

Neste ambiente, os peixes ganham características peculiares às condições ambientais que povoam: o *Anoplaster cornuta* possui dentes proeminentes e muito afiados que lhe permitem devorar qualquer criatura que passe na sua frente, pois, devido a escassez de comida, é importante aproveitar todas as oportunidades [3]; outras espécies emanam uma luminosidade que é a única

[8] A luz somente consegue penetrar no mar até em torno dos 1.000m, sendo que a primeira cor a desaparecer é a vermelha (6m de profundidade) e a última a azul [2].

existente neste ambiente [3].

Apesar das regiões abissais parecerem inóspitas, há um ecossistema diferente em profundidades um pouco menores. As placas tectônicas da terra estão em constante movimento. Em alguns lugares elas se aproximam em outros se afastam. Em 1977, a bordo do submersível Alvin, o geólogo americano Robert Ballard [6] descobriu os minivulcões a 2.400m de profundidade, perto das Ilhas Galápagos no Oceano Pacífico. Este ambiente se reproduz em vários lugares onde as placas tectônicas se afastam [38]. A cadeia Juan de Fuca tem sido estudada pelo Woods Hole Oceanographic Institute [54] há alguns anos. Neste ecossistema ocorre a mistura de água com ácido sulfídrico e outras substâncias que saem das chaminés, em torno de 400°C. Nesta profundidade, a pressão é maior que 200 atmosferas, sendo suficiente para esmagar uma baleia [9]. Na ausência de luminosidade e em uma temperatura em torno de 85°C vive o *Methanococcus jannaschii.* Neste ambiente aparentemente hostil a qualquer forma de vida, o *Methanococcus jannaschii* não apenas está adaptado como produz metano. Por ser um organismo autotrófico, serve de base para a teia alimentar deste ambiente. Os principais seres vivos neste ecossistema são os vermes tubulares que têm tamanho em torno dos 3m e possuem uma cabeça vermelha retrátil. Existem também moluscos brancos que podem chegar a 40cm, anêmonas, camarões, bactérias (arquebactérias) e visitantes eventuais como um tipo de polvo e o peixe *Pinkish eelpout* que se parece com uma cobra.

Além do aspecto biológico, há o aspecto geológico do ecossistema. Os minivulcões, também chamados de chaminés esbranquiçadas, jogam no mar uma mistura de água e várias substâncias inorgânicas: são os ventos hidrotermais. Eles fornecem o calor e as substâncias químicas para que os organismos autotróficos sirvam de base para a teia alimentar deste ecossistema. As chaminés são uma forma geológica muito instável: surgem e desaparecem num espaço ínfimo de tempo (geológico).

Apesar de toda esta adversidade do ambiente, o DNA do *Methanococcus jannaschii* apresenta as mesmas características do DNA de um ser qualquer da superfície da Terra: os mesmos 4 nucleotídeos, os mesmos 20 aminoácidos e o mesmo processo de síntese protéica. A maioria de suas proteínas ligadas à produção de energia, metabolismo e divisão celular são parecidas com as dos seres procarióticos; já as envolvidas nos processos de transcrição, tradução e replicação são em sua maioria similares às dos seres eucarióticos [53].

Nem toda a universalidade do código genético facilitou sua previsibilidade, sendo ainda um desafio a correta estimativa da entropia do DNA.

Capítulo 2

Tópicos da Biologia Molecular

A biologia molecular é uma área bastante ampla e complexa. Dentre os vários assuntos tratados por ela está o DNA. Para que haja uma compreensão básica do DNA, existente em todas as células, será feita uma revisão rápida sobre as células e seus constituintes moleculares. Este estudo está direcionado às proteínas e ácidos nucléicos com destaque para a estrutura do DNA e demais assuntos ligados aos artigos científicos tratados neste texto. Os aspectos funcionais da célula não serão detalhados. Este capítulo foi baseado em dois livros: um de biologia molecular de Zaha, et al. [58] e outro, didático, de biologia do ensino médio de Gewandsznajder e Linhares [22], tendo como bibliografia complementar as obras de Meidanis e Setubal [36], Nevill-Manning e Witten [39], Gusfield [26], Drlica [16] e Calladine e Drew [8].

2.1 Célula e seus constituintes moleculares

Em 1665 Robert Hooke observou que os tecidos de plantas eram divididos em compartimentos que ele chamou de *células*. Por volta de 1840 Theodor Schwann propôs que todos os seres vivos eram células ou agregados de células. Bem mais tarde, foi observado que haveria basicamente dois tipos de células e que os seres vivos podiam ser classificados em função do tipo de células que os compõe. Os organismos unicelulares como as bactérias, arquebactérias e algas azuis foram denominados *procarióticos*. Receberam a classificação de *eucarióticos* os seres de diversos tipos como protozoários (que são unicelulares), algas verdes, fungos, etc.

Nas células dos seres procarióticos não existe o núcleo celular; eles possuem um nucleóide sem a existência da membrana nuclear. Os organismos eucarióticos podem conter diversos tipos de células além de um núcleo bem definido e revestido pela membrana nuclear. É no núcleo que fica retido o DNA, exceto durante o processo da divisão celular. *Organelas* são regiões definidas dentro das células; são separadas do citoplasma por membranas internas e realizam funções específicas. Os seres procarióticos raramente possuem organelas; nos eucarióticos há

diversas organelas como mitocôndria, retículo endoplasmático, complexo de Golgi, lisossomo, cloroplasto, vacúolo etc.

As células são compostas por água, íons e outras substâncias; a água é o componente fundamental encontrado em maior quantidade nas células. Além de solvente natural para íons e minerais, é indispensável para os processos fisiológicos. Os íons servem para manter o equilíbrio ácido-básico da célula e também atuam na pressão osmótica. Há diversas outras substâncias químicas nas células, como as *pequenas moléculas* formadas por um número pequeno de átomos (em torno de 50) e que possuem características estruturais peculiares. Destaque para os *monômeros (ou resíduos)*: pequenas moléculas que servirão de base para a formação dos polímeros biológicos. Os *polímeros biológicos* são macromoléculas compostas por muitas cópias de um tipo de pequenas moléculas formando uma cadeia por ligações covalentes. Os nucleotídeos, aminoácidos, açúcares e ácidos graxos são os principais tipos de pequenas moléculas existentes.

Os *monossacarídeos* são moléculas pequenas e monoméricas, também chamadas de açúcares. Os *sacarídeos* ou *carboidratos* são compostos por uma ou mais moléculas de monossacarídeos. Dentre eles podemos citar os *oligossacarídeos* como a maltose, formados por poucas unidades de monossacarídeos, e os *polissacarídeos* como o amido e o glicogênio, compostos por diversos açúcares.

O *ácido graxo* é o lipídeo mais simples e entra na composição de lipídeos mais complexos como o glicerol, a gordura, e os fosfolipídeos. Estes são compostos por glicerol e ácido graxo e servem basicamente para formar as membranas celulares. Os ácidos graxos normalmente não são solúveis em água e dentre as suas funções está a composição da bicamada que constitui a membrana celular.

Os *polímeros biológicos* são compostos pelos monômeros. Existem quatro tipos básicos de monômeros e três tipos de polímeros biológicos: os polissacarídeos, as proteínas e os ácidos nucléicos; seus monômeros são respectivamente os açúcares, os aminoácidos e os nucleotídeos. Os ácidos graxos (monômeros) não possuem polímeros biológicos a eles associados.

2.2 Ácidos Nucléicos

Ácidos nucléicos são macromoléculas fundamentais aos seres vivos pois são essenciais ao processo da síntese protéica. São compostos por *nucleotídeos*: monômeros formados por três tipos de substâncias químicas: bases nitrogenadas, pentose e fosfato. As principais bases nitrogenadas são: adenina, guanina, citosina, timina e uracil. As bases adenina e guanina são conhecidas como *bases purínicas* ou *bases púricas*, derivadas da purina. As outras três bases são derivadas de um outro composto chamado pirimidina e são chamadas de *bases pirimidínicas* ou *bases pirimídicas*. As pentoses são de dois tipos: ribose ou desoxirribose. Há dois tipos de ácidos nucléicos: o ácido

desoxirribonucléico e o ácido ribonucléico. As bases adenina, citosina e guanina ocorrem tanto no ácido ribonucléico quanto no ácido desoxirribonucléico; a base uracil não aparece no ácido desoxirribonucléico bem como a timina não existe no ácido ribonucléico.

Polinucleotídeos são longos filamentos formados pela união dos nucleotídeos, sempre feita entre o fosfato de uma unidade e a pentose da unidade vizinha. Para simplificar a representação dos polinucleotídeos, convencionou-se a utilização das letras `'a'`, `'c'`, `'g'`, `'t'` e `'u'` para representar as bases adenina, citosina, guanina, timina e uracil, respectivamente. Embora haja uma ordem rígida para a ligação entre as duas cadeias polinucleotídicas, o mesmo não ocorre na ligação entre as bases de uma mesma cadeia ou *fita*.

2.3 Ácido ribonucléico

Ácido ribonucléico (RNA) é uma molécula de ácido nucléico formada por uma única cadeia polinucleotídica. Sua pentose é a ribose. Há diversos tipos de RNA, todos envolvidos de alguma forma na síntese protéica. Três deles merecem destaque: RNA mensageiro, RNA transportador e RNA ribossômico.

O *RNA mensageiro* (mRNA) contém a informação genética para a seqüência de aminoácidos. O *RNA transportador* (tRNA) transporta as moléculas de aminoácido até os ribossomos[1]. O *RNA ribossômico* (rRNA), que possui em torno de metade da massa dos ribossomos, está diretamente envolvido na síntese protéica. *Tradução* é o processo que converte a mensagem contida nas seqüências do RNA mensageiro em proteína. Foi demonstrado que o RNA ribossômico interage com o RNA mensageiro e os RNAs transportadores em diferentes estágios da tradução.

O processo de sintetizar o RNA mensageiro é chamado de *transcrição*. A molécula de RNA sintetizada é complementar à cadeia de DNA que lhe deu origem e idêntica à cadeia oposta, exceto pela presença de uracil no lugar da timina. Por convenção, a representação do DNA é sempre feita pela seqüência ou cadeia que chamamos de *codificadora* , que é a fita igual ao RNA que foi sintetizado. A cadeia que deu origem ao RNA por complementaridade não é representada. Para que haja a transcrição, é necessário que a dupla ligação do DNA seja aberta, para expor a cadeia que servirá de base para o RNA. Este processo é chamado de *desnaturação*. O processo contrário da religação das fitas do DNA é chamado de *renaturação*.

O *RNA polimerase* (RNAP) é uma enzima que possui múltiplas atividades: reconhece e liga-se a seqüências específicas do DNA, desnatura o DNA expondo a seqüência de nucleotídeos a ser copiada e mantém as fitas de DNA separadas na região em que estiver ocorrendo a transcrição. Também faz a renaturação do DNA na região que já ocorreu a síntese e termina a síntese só ou

[1] Presentes em todos os seres vivos, os ribossomos são formados por RNA e proteínas. Visíveis somente ao microscópio eletrônico, é nos ribossomos que ocorre a síntese protéica [22].

com o auxílio de outras proteínas específicas.

2.4 Ácido desoxirribonucléico

O DNA, abreviação em inglês para *ácido desoxirribonucléico* foi descoberto em 1869 por Friedrich Miesher (1844-1895). Sua pentose é a desoxirribose e ele é composto por nucleotídeos. Estudos de Chargaff entre 1949 e 1953 mostraram que a quantidade de bases adenina se iguala à quantidade de bases timina; o mesmo ocorre em relação à citosina e à guanina. Desta forma, o número de purinas é igual ao de pirimidinas. Por outro lado, a razão entre a quantidade de pirimidinas e de purinas varia consideravelmente entre as espécies.

Segundo o modelo de Watson e Crick de 1953, a molécula do DNA é constituída por duas cadeias polinucleotídicas dispostas em hélice ao redor de um eixo imaginário. As cadeias polinucleotídicas ficam unidas por meio de pontes de hidrogênio, que são mantidas entre bases específicas: a adenina com a timina e a citosina com a guanina. Estas duas cadeias formam uma estrutura em espiral, de dupla hélice, girando para a direita.

As cadeias do DNA são complementares. Devido à estreita forma de ligação nas moléculas do DNA, se em uma determinada posição de uma das duas cadeias de polinucleotídeos houver uma adenina, haverá necessariamente uma timina na posição equivalente da outra cadeia e vice-versa. O mesmo ocorre em relação às bases citosina e guanina. O fato das cadeias do DNA serem complementares simplifica a forma de representar a molécula de DNA; note-se que a representação de uma das cadeias indiretamente espelha a da outra cadeia.

Há três formas principais de representação do DNA. A *estrutura primária* é a seqüência das bases determinada por sua cadeia polinucleotídica . A *estrutura secundária* do DNA trata dos tipos de helicoidais formados. A *estrutura terciária* do DNA está ligada ao seu superenrolamento. Este trabalho lida exclusivamente com a representação da estrutura primária do DNA.

Uma das principais funções do DNA é guardar os códigos necessários para a sintetização de proteínas. Para a compreensão deste processo há vários conceitos que devem estar claros. Os *genes* são as regiões contíguas do DNA que contêm as informações necessárias para a produção das proteínas. São também as unidades básicas da hereditariedade. Os seres eucarióticos possuem seus genes divididos em duas regiões que se intercalam sucessivas vezes: são os íntrons e os exons. O conceito de gene foi proposto por Beadle e Tatum na década de 40: é a região do DNA responsável pela produção de uma proteína. Em 1956 Vernon Ingram demonstrou que as alterações no DNA produziam modificações na seqüência de aminoácidos da proteína específica. Os *íntrons* são as regiões dos genes que não estão ligadas à produção de proteínas e suas funções ainda não são conhecidas. Eles são retirados do mRNA após completada a transcrição. Os *exons* são as regiões dos genes que contêm as informações que são utilizadas após a transcrição

do mRNA. Em média, o tamanho dos íntrons é bem maior que o dos exons.

Em um organismo complexo como o organismo humano, as células são especializadas e não é toda célula que produz cada tipo de proteína. O processo proposto pelos cientistas franceses François Jacob, André Lwoff e Jacques Monod (Prêmio Nobel de 1965) prevê a existência de quatro tipos de genes ao longo da molécula de DNA: estrutural, regulador, promotor e operador. O gene estrutural é responsável pela transcrição de moléculas do mRNA. O gene promotor somente pode se ligar ao RNAP se seu vizinho gene operador estiver livre. Caso o operador esteja inibido, o promotor também o estará. Como o operador não se ligará ao RNAP, o processo de transcrição do mRNA não será ativado no gene estrutural. Toda esta ativação está ligada ao repressor, um tipo de proteína que é sintetizada a partir do gene regulador. Desta forma, de acordo com os genes ativados ou desativados em cada célula, a célula exerce uma função específica de seu tipo, permitindo assim a criação dos tecidos. No DNA, o espaço entre os genes é muito grande (bem maior que o tamanho do gene nos seres eucarióticos) e neste processo muitas vezes o gene regulador está bem distante do gene estrutural que está sendo inibido ou ativado.

2.5 Proteínas

As proteínas são polímeros que desempenham inúmeras funções biológicas e determinam a forma e a estrutura das células. São traduzidas a partir do mRNA, o qual foi transcrito de um segmento contíguo do DNA: o gene. Porém, ao contrário do DNA, os aminoácidos são a principal referência nas proteínas. Os *aminoácidos* são moléculas compostas por um átomo de carbono central (carbono α) com quatro ligações: uma para um átomo de hidrogênio (H), outra para o grupo amino (NH_2), outra para o grupo carboxilia ($COOH$) e a última para um radical que varia de aminoácido para aminoácido. O radical pode ser tão simples quanto um átomo de hidrogênio, como na glicina ou bem mais complexa, dependendo do aminoácido.

Os *códons* são seqüências de três nucleotídeos e são utilizados para representar os aminoácidos através do código genético. O *código genético* é composto por nucleotídeos. Ele é empregado para representar as proteínas e é *degenerado* ou *redundante*, pois há mais de um códon com o mesmo significado. Por exemplo, a alanina pode ser representada pelos códons: `gcu`, `gcc`, `gca` ou `gcg`. Como as duas primeiras bases destes códon não variam, é possível que no passado os aminoácidos fossem representados por apenas duas bases. A degeneração, de alguma forma, protege os seres vivos contra eventuais mutações genéticas, tornando-os mais estáveis.

Dos 64 possíveis códons distintos, 61 deles representam cada um dos 20 aminoácidos existentes. Os códons `uaa`, `uga` e `uag` não representam nenhum aminoácido em específico; representam o final de codificação. O códon `aug`, Metionina, é um dos sinais que determina o início da síntese

Código	aminoácido	trios de bases (códons) no DNA
A	Alanina	gcu, gcc, gca, gcg
C	Cisteína	ugu, ugc
D	Ácido Aspártico	gau, gac
E	Ácido Glutâmico	gaa, gag
F	Fenilalanina	uuu, uuc
G	Glicina	ggu, ggc, gga, ggg
H	Histidina	cau, cac
I	Isoleucina	auu, auc, aua
K	Lisina	aaa, aag
L	Leucina	uua, uug, cuu, cuc, cua, cug
M	Metionina	aug
N	Asparagina	aau, aac
P	Prolina	ccu, ccc, cca, ccg
Q	Glutamina	caa, cag
R	Arginina	cgu, cgc, cga, cgg, aga, agg
S	Serina	agu, agc, ucu, ucc, uca, ucg
T	Teonina	acu, acc, aca, acg
V	Valina	guu, guc, gua, gug
W	Triptófano	ugg
Y	Tirosina	uau e uac

protéica. É o primeiro aminoácido a ser incorporado, tanto nas proteínas de procarióticos como nas de eucarióticos, com raras exceções. Note-se que há vários aminoácidos com mais de um códon que o represente.

As proteínas representam mais da metade do peso seco (descontado o peso da água) de uma célula e são conhecidas como as moléculas que realizam o trabalho celular. A ligação dos aminoácidos formando uma proteína é chamada de *ligação peptídica*. A *estrutura primária* de uma proteína é a seqüência de aminoácidos que formam a sua cadeia peptídica. A *estrutura secundária* lida com a organização espacial dos aminoácidos próximos na cadeia peptídica. A *estrutura terciária* se refere à forma que a cadeia polipeptídica está enovelada. Finalmente a *estrutura quaternária* trata da disposição das subunidades protéicas que formam a molécula.

2.6 Um pouco mais sobre o DNA

Alguns algoritmos exploraram uma característica comum no DNA descrita por Zaha et al. [58], as *seqüências complementares invertidas* ou *estruturas cruciformes* ou *repetições invertidas*.

"Existem seqüências no DNA que têm simetria, contendo seqüências repetidas invertidas. Por exemplo:

attcgcgtagtagacatagctgacatagtcagctatgtctgctc

Seqüências como essa podem formar estruturas onde o pareamento entre duas fitas é substituído pelo pareamento entre as bases complementares na mesma fita. Tal alteração reduz o número de voltas da dupla-hélice na região, removendo o superenrolamento negativo, um para cada dez bases de pareamento. Essas estruturas podem ser formadas *in vivo*."

Zaha et al. [58] também descrevem uma teoria sobre a origem dos íntrons. A origem dos genes contendo íntrons está relacionada ao fato de quase a totalidade dos seres eucarióticos possuírem este tipo de gene, enquanto com os seres procarióticos ocorre exatamente o contrário. Há evidências que sugerem que os genes primitivos eram constituídos de pequenos pedaços de DNA que codificavam proteínas simples. Dois genes primitivos poderiam fundir-se para formar uma proteína mais complexa. Os íntrons permitem o rearranjo de seqüências codificantes, sem a necessidade de justaposição, mas impõem a necessidade de um mRNA que permita a retirada dos íntrons antes da síntese, a teoria de *exon shuffling*.

Esta teoria é baseada na evidência atual que exons, ou grupos de exons, freqüentemente codificam domínios específicos em algumas proteínas, domínios estes funcionalmente diferentes. Outra evidência é que genes diferentes possuem alguns exons iguais entre si, mas não todos, sugerindo que esses genes foram originados por adição de exons como se fossem módulos.

Processo de formação do DNA

O DNA possui informação e a mensagem é um meio para permitir o transporte da informação. O transporte de informação na célula implica a existência de um código comum ao transmissor e ao receptor, o código genético. Ele é conhecido há décadas: os nucleotídeos codificam a informação e esta é passada para os processos de reprodução, divisão celular e síntese de proteínas. Desta forma, a partir de uma única célula, que é o zigoto, um novo organismo é criado no processo de reprodução. A informação sai do DNA e passa ao zigoto induzindo que este seja dividida um certo número de vezes; depois uma certa quantidade de células formam o coração, outra o cérebro e assim para cada órgão e tecido.

O código genético é conhecido, mas não o é a forma pela qual o código genético representa a vida. Os exons servem para codificar as proteínas. A finalidade dos íntrons e do grande espaçamento existente entre os genes (bem maior que os próprios genes) é ignorada.

O estudo do código genético somente foi possível depois que Pasteur, no século XIX, provou que a vida não é gerada de forma espontânea [20]. Também no século XIX Mendel deduziu os princípios básicos da hereditariedade. Somente na década de 20 (século XX) houve a descoberta de que a maioria das características são definidas por dúzias de genes [20]. Após a segunda guerra mundial, utilizando os avanços tecnológicos do pós guerra, foi possível à biologia molecular chegar à sua forma atual.

O conhecimento mais importante que liga o código genético às cadeias de DNA é o processo evolutivo. Se a mutação genética for vantajosa, ela será passada aos descendentes do indivíduo e será incorporada à espécie, se houver tempo suficiente.

A seleção natural está vinculada ao meio que a espécie vive (hábitat). Existe uma grande quantidade de ecossistemas; dos ventos hidrotermais ao deserto do Saara. Indivíduos da mesma espécie, em ambientes diferentes podem sofrer processos distintos de seleção natural. Em qualquer espécie, são produzidos mais indivíduos do que os que reproduzem. Na competição entre indivíduos da mesma espécie, somente os mais adaptados conseguem se reproduzir. A mutação genética será vantajosa se gerar para o indivíduo uma melhor adaptação ao hábitat. Há diversas implicações neste processo, como as ondas de evolução, por exemplo, que não fazem parte do escopo deste trabalho. Somente para as proteínas é conhecido o processo pelo qual a mutação genética é responsável por uma melhor adaptação dos indivíduos ao hábitat.

Embora a seleção natural e a teoria da evolução não sejam aceitas universalmente é o único fator conhecido no processo de formação do DNA (como fonte de informação). Talvez este seja o grande problema na previsibilidade do DNA: o desconhecimento de seus processos de formação.

CAPÍTULO 3

Teoria da Informação e Suas Bases Probabilísticas

Uma das características da informação é sua possibilidade de ser transmitida. A informação é expressa de alguma forma. Esta expressão pode ser transmitida e também estudada. No contexto da computação, a expressão da informação é freqüentemente chamada de *mensagem*, que é representada normalmente por uma cadeia de símbolos. Uma mensagem pode ser estudada em pelo menos três aspectos distintos:

- *Informação sintática* destaca os símbolos de uma mensagem e as diversas correlações entre eles.
- *Informação semântica* enfatiza o significado da mensagem.
- *Informação pragmática* relaciona-se ao uso ou efeito da mensagem.

Uma forma de mostrar a diferença das abordagens é através de um exemplo. Considere a mensagem

``ggatccagattcttttgaaattcctcctgc...''

No enfoque sintático, um estudo desta seqüência pode medir, por exemplo, o número de ocorrências do símbolo a ou a quantidade de vezes que o símbolo t é precedido pelo símbolo t. No contexto semântico, esta mensagem mostra os trinta primeiros nucleotídeos do gene Humretblas. O aspecto pragmático considera a maior ou menor susceptibilidade ao retinoblastoma de uma pessoa com esta cadeia genética.

Neste texto, o termo 'informação' se refere à informação sintática, a menos que seja dito o contrário. A informação fica reduzida ao estudo de sua forma de expressão, deixando de lado seus demais aspectos. Nas aplicações da teoria da informação à biologia molecular, é feito um

estudo sintático da representação das cadeias de nucleotídeos ou aminoácidos que formam algum trecho de cadeia de DNA ou proteína. A partir deste estudo da informação sintática, existe a possibilidade de inferências sobre a semântica dessas cadeias, ou seja, sobre seus aspectos biológicos. Outra abordagem que tem sido utilizada é o estudo sintático das cadeias, cujo significado semântico é conhecido, onde são feitas medições na busca de padrões para estas informações semânticas.

Este trabalho usa expressões técnicas simplificadas: "a entropia do DNA" deve ser entendida como "o cálculo da entropia feito na representação da fita codificadora do DNA através da representação dos nucleotídeos que a formam"; "a compressão de proteínas" se refere a "um algoritmo de compressão utilizado na representação da proteína através da cadeia de aminoácidos que a formam ou da cadeia polinucleotídica que a codifica". O rigor dos significados obrigaria a utilização da forma completa. Mas nesta dissertação, a representação simplificada deve ser interpretada como uma forma de representação da expressão completa. Este trabalho buscou a análise matemática do aspecto sintático da informação contida nas cadeias genômicas e não contou com o auxílio de biólogos.

A entropia está ligada à teoria da informação que por sua vez é responsável por uma boa parte dos conceitos básicos da compressão de dados. Como alguns artigos vistos utilizam as técnicas de compressão na avaliação das estimativas de entropia do DNA, os conceitos ligados à teoria da informação enfocarão a compressão de dados. A compressão de dados foi desenvolvida a partir dos estudos feitos para compressão de textos. Foram fontes para os conceitos da teoria da informação os livros de Bell, Cleary e Witten [5], de Sayood [47] e da escola européia de teoria da informação van der Lubbe [52].[1] Em alguns artigos há complementos desta teoria mas estes estão restritos aos escopos dos mesmos.

Neste capítulo será apresentado um esboço da definição de modelo. Ela será detalhada no capítulo seguinte, pois é extremamente relevante e necessária para a estimativa da entropia e amplamente utilizada nas técnicas de compressão.

A teoria da informação, modelos e compressão utilizam conceitos de probabilidade. Foi feita uma rápida revisão destes conceitos através dos livros de Meyer [37] e de Cormen, Leiserson e Rivest [11].

[1] Deste livro de teoria da informação foram utilizados aspectos mais gerais, pois há visíveis diferenças entre as escolas européias e americanas no trato da teoria da informação. Apesar da citação do artigo de Shannon que definiu a entropia [49], o livro não contém o termo entropia nem em seu índice remissivo. Ele manteve o aspecto semântico e pragmático da escola européia, com abordagem reduzida sobre o aspecto sintático da escola americana. Em tempo, a entropia é vista aí como a medida da quantidade de informação de Shannon.

3.1 Conceitos de probabilidade

A probabilidade utiliza com muita freqüência experimentos aleatórios. *Experimentos aleatórios* ou *não-determinísticos* são aqueles que podem ser repetidos indefinidamente sob as mesmas condições (inalteradas). Eles possuem um conjunto de resultados possíveis que sempre pode ser definido. Se um experimento for repetido um número suficientemente grande de vezes, ele apresentará uma regularidade nos resultados. Exemplos de experimentos aleatórios são: as temperaturas mínimas e máximas registradas no período de um dia em um determinado local, o lançamento de um dado e seu resultado e jogar uma moeda 4 vezes verificando o número de caras. O *espaço amostral* de um experimento é o conjunto de todos os resultados possíveis. Para os exemplos acima, os espaços amostrais são respectivamente $\{(x_{\min}, x_{\max}) \in \mathbb{R}^2 : -70 \leq x_{\min} < x_{\max} \leq 70\}$; $\{1,2,3,4,5,6\}$ e $\{0,1,2,3,4\}$. Se Ω é um espaço amostral, então um *evento*[2] é qualquer subconjunto de Ω. Exemplos de eventos para os espaços amostrais acima são: $\{x_{\min} = 18, x_{\max} = 25\}$ (temperatura mínima 18 e máxima 25); $\{4,5,6\}$ (sair um número maior que 3 no lançamento de um dado) e $\{2\}$ (ocorrerem exatamente duas caras ao se jogar uma moeda quatro vezes). A *freqüência relativa* de um evento A para um experimento repetido n vezes é $f_A = \frac{n_A}{n}$, sendo n_A o número de vezes que A ocorreu, ou seja, o número de vezes em que foi obtido algum elemento do evento A ao se repetir o experimento n vezes. Note que f_A é um número entre 0 e 1. Se A representa a totalidade do espaço amostral então $f_A = 1$. Podem ocorrer eventos com probabilidade nula; no exemplo das temperaturas nunca ocorrerão as temperaturas $x_{\min} = -70$ e $x_{\max} = 70$ no mesmo dia e local.

A *probabilidade* de um evento A é a freqüência relativa deste evento quando n tende a ∞.

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

Os eventos $A_1, \ldots, A_k$ sobre um experimento formam uma *partição do espaço amostral* se $P(A_i) > 0$ para cada i, A_i e A_j são disjuntos para cada i diferente de j e se $A_1 \cup \ldots \cup A_k$ for a totalidade do espaço amostral. $P = \{P_1, \ldots, P_n\}$ será uma *distribuição discreta de probabilidades* se $P_i \geq 0$ para todo i e $\sum_{i=1}^{n} P_i = 1$. Sejam A e B eventos de algum experimento. Se $P(A \cap B) = P(A) \times P(B)$ então A e B são ditos *eventos independentes*. A *probabilidade condicional* $P(A|B)$ de ocorrer o evento A dado que ocorreu o evento B é dada por $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Os eventos $A_1, \ldots, A_k$ de algum espaço amostral são chamados de *mutuamente independentes* se $P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1) \times P(A_2) \times \ldots \times P(A_k)$.

Uma *variável aleatória* X é uma função que associa um número real a cada elemento do espaço amostral Ω. Nesta dissertação, quando houver alguma referência à probabilidade de uma variável aleatória, ela será representada através do valor por ela assumido. Um exemplo

[2] Eventos e *eventos aleatórios* possuem o mesmo significado para este texto. Quando aparecer, por exemplo, "Sejam A e B eventos aleatórios" seu significado é: dado um experimento aleatório que possua seu espaço amostral, A e B são subconjuntos deste espaço amostral.

desta representação é: $P(a) = P(X = a)$, é a representação da probabilidade da variável aleatória X assumir o valor a. O *valor esperado* ou *esperança matemática* $E(X)$ de uma variável aleatória X é a média ponderada dos valores possíveis, se estes forem finitos, sendo definida matematicamente por

$$E(X) = \sum_{x \in X(\Omega)} xP(x)$$

A esperança matemática é bem definida se a soma for finita ou convergir absolutamente (não ser ∞). A *variância* $V(X)$ de uma variável aleatória X é definida por $V(X) = E(X^2) - [E(X)]^2$ e o desvio padrão $\sigma(X)$ é a raiz quadrada da variância de X. Logo, $V(X) = \sigma^2(X)$.

Seja Ω um espaço amostral de algum experimento. Considera-se que o par $(X(\omega), Y(\omega))$, ω em Ω, é uma *variável aleatória bidimensional*. Um exemplo está nas temperaturas mínimas e máximas de um certo local no período de um dia. Sejam $X = \{x_1, \ldots, x_n\}$ o conjunto de todas as temperaturas mínimas possíveis e $Y = \{y_1, \ldots, y_m\}$ o conjunto de todas máximas possíveis. A probabilidade conjunta $P(x_i, y_j)$ é a probabilidade da temperatura mínima ser $x_{\min} = x_i$ e a temperatura máxima $x_{\max} = y_j$. Nestas circunstâncias também pode haver uma distribuição discreta de probabilidades se $P(x_i, y_j) \geq 0$ para cada x_i e y_j e $\sum_{x_{\min} \in X} \sum_{x_{\max} \in Y} P(x_{\min}, x_{\max}) = 1$. É possível estender este processo para variáveis n-dimensionais.

Os conceitos de probabilidade vão muito além dos aqui mostrados. Estes podem ser expandidos para o domínio da matemática contínua. Porém, a utilização da probabilidade nas cadeias genômicas se restringe à matemática discreta. Mesmo assim, muitas das propriedades da probabilidade na matemática discreta são inúteis para as aplicações na biologia molecular.

3.2 Modelos

Pode ocorrer de alguma mensagem ser gerada a partir de algum método específico, o *modelo gerador*. As mensagens criadas desta forma refletem o modelo gerador se forem suficientemente grandes. Um exemplo de modelo gerador é a cadeia de Markov de tempo discreto que será vista a seguir.

Processo de Markov

As cadeias filogenéticas que relacionam uma proteína a outra através de mutações são exemplos de transmissão entre estados de um sistema [56]. Na biologia molecular existe algum interesse nos chamados processos de Markov discretos e finitos. São ferramentas que auxiliam na determinação deste relacionamento entre as proteínas.

Um dos modelos muito utilizados na teoria da informação é um tipo específico de processo de

Markov[3], chamado *cadeia de Markov de tempo discreto* [47] ou *cadeia de Markov estacionária.* Seja Σ um alfabeto e $s_1 \ldots s_n$ uma cadeia com n elementos de Σ. Esta cadeia é dita uma cadeia de i-ésima ordem do modelo de Markov se i for o menor valor para o qual a expressão abaixo seja verdadeira:

$$P(s_k = \sigma_k | s_{k-1} = \sigma_{k-1}, \ldots, s_{k-i} = \sigma_{k-i}) = P(s_k = \sigma_k | s_{k-1} = \sigma_{k-1}, \ldots, s_{k-i} = \sigma_{k-i}, \ldots),$$

onde σ são símbolos em Σ. O conhecimento dos i antecessores imediatos equivale ao conhecimento de todos os antecessores para a previsibilidade do próximo símbolo. Os valores tomados do conjunto $\{\sigma_{k-1}, \ldots, \sigma_{k-i}\}$ são chamados *estados do processo.* Sendo λ o tamanho do alfabeto, o número de estados do processo é λ^i, ou seja, a quantidade de estados no processo de Markov cresce exponencialmente à ordem do processo.

Este processo é *estacionário* porque a previsibilidade do próximo símbolo depende do estado mas não da posição deste na cadeia. O processo é dito *ergódico* (ou a cadeia gerada por ele) se a probabilidade de transição de um estado a outro do processo é positiva, para todo par de estados.

Classes de condição

Freqüentemente, um evento é representado por uma seqüência de símbolos sobre um alfabeto Σ. Seja $\sigma_1\sigma_2 \ldots \sigma_n$ uma cadeia sobre Σ. A *classe de condição* da cadeia $\sigma_j\sigma_{j+1} \ldots \sigma_k$ é o conjunto das probabilidades condicionais tais que o evento $\sigma_j\sigma_{j+1} \ldots \sigma_k$ ocorreu. Em um contexto mais amplo, a seqüência de símbolos remete ao início da mensagem, mas uma aproximação válida é calcular a probabilidade condicional para um número limitado de símbolos (dois símbolos por exemplo).

Cada estado de um processo de Markov pode ser associado a uma classe de condição.

3.3 Compressão

Uma vez que assumiu alguma forma de representação a informação pode ser transmitida. A *codificação* é o processo de modificar a forma de representação da mensagem, sem lhe alterar o conteúdo. A *decodificação* é o processo contrário de representar a mensagem no original de sua representação. O *código* é um meio de converter alguma forma de expressão da mensagem em outra. Um exemplo bem simples está na representação de textos escritos num computador. O bit, unidade de representação da memória, somente reconhece o pulso elétrico e pode estar ligado `1` ou desligado `0`. Como o computador não reconhece as letras do texto escrito, foram criados vários

[3] Andrei Andrevich Markov foi um matemático russo que viveu de 1856 a 1922.

códigos para representá-las. Para o EBCDIC a representação do espaço é `00101000`. Já no código ASCII a representação do espaço é `00100000`. Dentro do contexto da computação, há um custo envolvido no processo de transmissão. Para diminuir este custo é feito um processo especial de codificação chamado compressão. *Compressão* é a tentativa de reescrever a mensagem utilizando menos espaço. O processo de comprimir também é chamado de codificação; já o processo inverso é a decodificação. O *transmissor* é aquele que efetiva a codificação ou compressão e o *receptor* o que faz a decodificação.

Um *alfabeto* é um conjunto finito de símbolos possíveis de alguma mensagem. A mensagem é o alvo da compressão. Um conceito restritivo de alfabeto é aquele em que cada símbolo representa exatamente um caractere (equivalente a uma partição do espaço amostral no qual cada evento é um caractere possível de ocorrer no experimento caracteres de uma seqüência). Generalizando, o alfabeto é um conjunto finito dos símbolos, (um ou mais caracteres) que não são formados por uma seqüência qualquer de símbolos do alfabeto (equivalente a uma partição qualquer do espaço amostral no qual cada evento é uma seqüência não vazia de caracteres que podem ocorrer na mensagem).

A mensagem guarda uma certa *quantidade de informação*. Esta não está diretamente relacionada ao tamanho da mensagem. Em um exemplo, números naturais consecutivos podem ser representados de diversas formas distintas. A representação de todos os números naturais de um a um milhão pode utilizar 5.888.896 algarismos $= (9 \times 1) + (90 \times 2) + (900 \times 3) + (9.000 \times 4) + (90.000 \times 5) + (900.000 \times 6) + (1 \times 7)$ (ao escrever cada número) ou com outra representação na qual são empregados os 44 caracteres da mensagem ``Todos os números naturais de um a um milhão.'' ou na representação matemática, mais compactada: $1, 2, \ldots, 1.000.000$ com exatos 17 caracteres. Neste exemplo a quantidade de informação é a mesma, embora o tamanho das mensagens que representam a informação varie de 17 a 5.888.896 caracteres.

A previsibilidade da mensagem interfere diretamente na eficiência de sua compressão. Quanto mais previsível for a mensagem melhor será a sua compressibilidade. O *modelo* é um meio utilizado para definir uma distribuição de probabilidades da mensagem (sua previsibilidade). Se N_i é o tamanho da mensagem antes da compressão e N_f o tamanho da mensagem comprimida, então a *taxa de compressão* τ é a razão

$$\tau = \frac{N_f}{N_i}.$$

Sayood [47] utiliza

$$\tau = 1 - \frac{N_f}{N_i}$$

para definir a taxa de compressão. Desta forma ela estará referindo à redução sofrida pela mensagem ao ser comprimida e será uma taxa negativa se a mensagem comprimida for maior que a mensagem original. A taxa de compressão depende da quantidade de informação. A quantidade

de informação depende da previsibilidade da mensagem. Quanto maior for a previsibilidade da mensagem menor será a quantidade de informação nela contida. Já a previsibilidade da mensagem depende do modelo utilizado para gerar a distribuição de probabilidades. Quanto menos uniforme for a distribuição de probabilidades, mais previsível é a mensagem.

O processo de compressão pode ser reversível ou irreversível. A compressão *reversível* (sem perdas) é aquela que permite recuperar a mensagem original integralmente. É o tipo de compressão normalmente utilizada para textos. A compressão *irreversível* é aquela que não permite a recuperação integral da mensagem original, como acontece na conversa telefônica. Na telefonia há pequenas alterações na voz dos interlocutores, mas é perfeitamente possível distinguir quem está falando e o que está sendo dito. Os processos de compressão utilizados nas cadeias de DNA são sempre reversíveis.

3.4 Entropia

A entropia foi definida em 1864 por Rudolf Clausius para a Física e em 1949 por Shannon para a teoria da informação [17].

A entropia é um número vinculado a uma distribuição de probabilidades. É possível calcular a entropia de uma distribuição de probabilidades fora do contexto da teoria da informação, porém este caso não é abordado neste texto. No contexto da teoria da informação, quando é calculado algum valor para a entropia, está se criando um vínculo entre uma seqüência, um modelo e a entropia da distribuição de probabilidades fornecido pela aplicação do modelo à seqüência. Farach, et al. [19] mostram a entropia vinculada ao número de entradas no dicionário. Este tipo de abordagem foi desenvolvido por Ziv e Lempel [59] e é mostrada no capítulo 6.

O valor da estimativa da entropia é obtido através de cálculos sobre a distribuição de probabilidades fornecida pela aplicação de algum modelo à mensagem. Porém, quando se conhece o modelo gerador de uma mensagem, é possível calcular sua entropia de uma forma mais simples.

Entropia de uma mensagem

A *entropia de uma mensagem obtida a partir de um modelo gerador* é função de sua probabilidade de ser gerada por este. Se M é o modelo gerador, S é alguma mensagem obtida através deste modelo e $P(S)$ é a probabilidade de S ser gerada por M então a entropia de $H(S)$ é definida por:

$$H(S) = -\log_2 P(S) \text{ bits.} \tag{3.1}$$

A probabilidade de uma mensagem obtida a partir de um modelo considera a probabilidade de cada símbolo na cadeia, o tamanho da cadeia e a probabilidade do símbolo de início que possui

um valor especial se o modelo gerador não for ergódico. Mensagens equiprováveis possuem o mesmo valor de entropia.

Entropia de uma distribuição de probabilidades

Informação e incerteza são termos que descrevem um processo que seleciona um ou mais objetos de um conjunto de objetos. A incerteza é medida através da função logaritmo. A base do logaritmo define a unidade da incerteza. Se for utilizada a base 10, a incerteza será medida em número de dígitos [48]. Ao se utilizar a base 2, em quantidade bits e esta será adotada como padrão ao longo deste texto. Seja $\Sigma = \{\sigma_1, \ldots, \sigma_\lambda\}$ um alfabeto e suponha os símbolos do alfabeto equiprováveis, nestas condições a *incerteza* U é definida por

$$U = \log_2 \lambda$$

Desta forma se um alfabeto possui apenas 1 símbolo, a incerteza é $\log_2 1 = 0$ bit. Trabalhando um pouco esta expressão obtém-se que

$$\begin{aligned} \log_2 \lambda &= -\log_2 (\lambda^{-1}) \\ &= -\log_2 \frac{1}{\lambda} \end{aligned}$$

sendo $1/\lambda$ a probabilidade da seleção de qualquer símbolo do alfabeto.

Ocorre que raras vezes há uma distribuição uniforme das probabilidades. A incerteza pode ser generalizada para uma distribuição qualquer de probabilidades dos λ símbolos no alfabeto Σ. Seja $P_1, P_2, \ldots, P_\lambda$ uma distribuição discreta de probabilidades, ou seja: $P_i \geq 0$ para todo i e

$$\sum_{i=1}^{\lambda} P_i = 1,$$

onde P_i é a probabilidade de σ_i ocorrer. A incerteza de um símbolo σ_i qualquer é

$$u_i = -\log_2 P_i$$

Eventualmente pode ser interessante calcular a incerteza de uma mensagem S com n símbolos. Seja n_i o número de ocorrências do símbolo σ_i na mensagem, tem-se que:

$$n = \sum_{i=1}^{\lambda} n_i$$

A *incerteza média* da mensagem será dada por

$$\frac{\sum_{i=1}^{\lambda} n_i u_i}{n}$$

sendo $\frac{n_i}{n}$ a probabilidade de ocorrer o símbolo σ_i e $u_i = -\log_2 P_i$. Trabalhando um pouco mais a expressão acima tem-se a incerteza média da mensagem.

$$\begin{aligned} & \frac{\sum_{i=1}^{\lambda} n_i u_i}{n} \\ = & \sum_{i=1}^{\lambda} P_i u_i \\ = & -\sum_{i=1}^{\lambda} P_i \log_2 P_i \\ = & \frac{1}{n} H(S). \end{aligned}$$

A incerteza média da mensagem está relacionada com a entropia da mensagem obtida por um modelo gerador (equação (3.1)), sendo expressada em bits por símbolo.

De uma maneira mais geral, Shannon [49] chamou de *entropia da distribuição de probabilidades* $P = \{P_1, P_2, \ldots\}$ a quantidade

$$H(P) = -\sum_i P_i \log_2 P_i. \tag{3.2}$$

A entropia é a medida da *quantidade de informação de Shannon* para a distribuição de probabilidades P [52]. Na teoria da informação, a distribuição de probabilidades está vinculada a uma mensagem ou a algum trecho de uma mensagem (seqüência). Se o cálculo da entropia for feito sobre a distribuição de probabilidades do conjunto de símbolos do alfabeto, sua dimensão será bits por símbolo.

Entropia condicional

Sejam X e Y variáveis aleatórias. Suponha que X assume valores sobre o alfabeto de origem[4] $\Sigma_X = \{x_1, \ldots, x_i\}$ e Y sobre o alfabeto de reconstrução[5] $\Sigma_Y = \{y_1, \ldots, y_j\}$. Seja ainda $P(x_i) = P(X = x_i)$ e $P(y_j) = P(Y = y_j)$. Foi visto nesta seção que a entropia da distribuição de probabilidades é dada pela equação de Shannon (3.2):

$$H(X) = -\sum_{k=1}^{i} P(x_k) \log_2 P(x_k) \text{ e } H(Y) = -\sum_{k=1}^{j} P(y_k) \log_2 P(y_k).$$

A *auto-informação* de um evento A é definida por:

$$i(A) = \log_2 \frac{1}{P(A)} = -\log_2 P(A)$$

[4] Alfabeto da mensagem original.

[5] Alfabeto da mensagem comprimida, a partir da qual será reconstruída a mensagem original pelo decodificador.

e a *auto-informação condicional* do evento A dado que ocorreu o evento[6] B é definida por:

$$i(A|B) = -\log_2 P(A|B)$$

sendo $P(A|B)$ a probabilidade condicional conforme visto na seção (3.1).

O valor médio da auto-informação condicional é chamado de *entropia condicional.* A entropia condicional $H(X|Y)$ pode ser interpretada como a quantidade de incerteza restante sobre a variável X, dada a reconstrução da variável Y. Matematicamente a entropia condicional $H(Y|X)$ é:

$$H(X|Y) = -\sum_{x\in\Sigma_X}\sum_{y\in\Sigma_Y} P(x|y)P(y)\log_2 P(x|y)$$

Teoricamente há a limitação $H(X|Y) \leq H(X)$ de forma que o conhecimento da variável Y reduz a incerteza em relação à variável X [47].

Entropia e informação mutual

Sejam X e Y variáveis aleatórias. Suponha que X assume valores sobre o alfabeto de origem $\Sigma_X = \{x_1, \ldots, x_l\}$ e Y sobre o alfabeto de reconstrução $\Sigma_Y = \{y_1, \ldots, y_m\}$. Seja ainda $P(x_k) = P(X = x_k)$ e $P(y_j) = P(Y = y_j)$. A informação mutual entre x_k e y_j é definida matematicamente como:

$$i(x_k; y_j) = \log_2 \frac{P(x_k|y_j)}{P(x_k)}.$$

A *entropia mutual* ou *informação mutual média* é o relacionamento que existe entre a entrada e a saída do canal [56] e é definida por:

$$\begin{aligned} I(X;Y) &= \sum_{k=1}^{l}\sum_{j=1}^{m} P(x_k, y_j)\log_2\left[\frac{P(x_k|y_j)}{P(x_k)}\right] \\ &= \sum_{k=1}^{l}\sum_{j=1}^{m} P(x_k, y_j)\log_2 P(x_k|y_j) - \sum_{k=1}^{l}\sum_{j=1}^{m} P(x_k, y_j)\log_2 P(x_k) \\ I(X;Y) &= H(X) - H(X|Y) \end{aligned}$$

A entropia mutual é simétrica com relação ao transmissor e ao receptor, desta forma $I(X;Y) = I(Y;X)$ e pode ser escrita como:

$$I(X;Y) = H(Y) - H(Y|X) = I(Y;X). \tag{3.3}$$

[6] Em particular, se $A_1 \ldots A_k$ são eventos independentes e $A_1 \cup \ldots \cup A_k = \Omega$, sendo Ω o espaço amostral, então $H = \sum P(A_i)i(A_i)$ é a entropia associada ao experimento.

Entropia do processo de Markov

O cálculo da entropia de uma cadeia que segue o processo de Markov é feito pelo produto da probabilidade de cada estado do processo pela entropia do estado. A entropia de cada estado reflete sua incerteza. Considere $H(e)$ a *entropia do estado*:

$$H(e) = -\sum_{\sigma \in \Sigma} P(\sigma | s_1 \dots s_k) \log_2 P(\sigma | s_1 \dots s_k) \tag{3.4}$$

$s_1 \dots s_k$ representa algum estado e. A entropia de cadeias que seguem alguma ordem do processo de Markov é calculada pela expressão:

$$H = \sum_e H(e) P(e) \tag{3.5}$$

onde $P(e)$ é a probabilidade de ocorrer o estado e na cadeia.

Entropia e compressão

Seja S é uma cadeia ergódica com tamanho n sobre um alfabeto Σ_1 e $P(S)$ a probabilidade de S. Se algum esquema de compressão for aplicado a S então haverá a cadeia codificada S', com o tamanho n' sobre um alfabeto Σ_2. A taxa de compressão de S é n'/n e o valor esperado para a taxa de compressão das cadeias ergódicas de tamanho n é definida como:

$$\tau_n = \frac{\sum_s P(S) n'}{n} \tag{3.6}$$

onde o somatório é tomado sobre todas as cadeias ergódicas de tamanho n. Quando n cresce o valor de τ_n se aproxima de τ, o coeficiente de compressão.

Shannon [49] mostrou que se H é a entropia da cadeia de Markov que gerou as seqüências do somatório em (3.6) então

$$\tau \geq H / \log |\Sigma_1 \cup \Sigma_2|.$$

Se $\Sigma_1 = \Sigma_2 = \{0, 1\}$ então

$$\tau \geq H$$

e existe τ' tal que

$$\tau' \leq H + 1.$$

O teorema da codificação sem interferência de Shannon mostra que o número médio de símbolos binários da mensagem comprimida pode se aproximar da entropia do modelo gerador mas não pode ser menor [5]. Contudo, existe uma restrição na quantidade de compressão que pode ser obtida nos esquemas reversíveis. O limite é definido pela entropia da mensagem original (antes da compressão), abaixo da qual é impossível reduzir o tamanho da seqüência comprimida [47].

Modelos

A tarefa fundamental de um modelo é fornecer a distribuição de probabilidades para uma mensagem. Muitas vezes as distribuições de probabilidades são calculadas de forma incremental, começando no início da mensagem, e 'adaptando' as probabilidades à medida que seus símbolos são examinados.

A distribuição de probabilidades obtidas a partir de um modelo fornece a previsibilidade de uma mensagem. Normalmente uma mensagem é escrita sobre um alfabeto Σ. O tamanho de uma mensagem (n) é o número de símbolos que nela ocorrem. Dado um alfabeto $\Sigma = \{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}, \texttt{e}\}$ a mensagem

$$S = \texttt{``eaebbdecbaedbdecaaedbaecddecbadd''}.$$

possui 32 símbolos.

Seja M_1 o modelo aplicado pelo processo de contagem das ocorrências de cada símbolo do alfabeto e obtenção da probabilidade de cada símbolo pela divisão do contador do símbolo pelo tamanho da mensagem. Seja $P(\texttt{x})$ a probabilidade de ocorrer o símbolo x em S. Utilizando o modelo M_1 para o cálculo da distribuição de probabilidades, esta ficou para S:

$$P(\texttt{a}) = 6/32;\ P(\texttt{b}) = 6/32;\ P(\texttt{c}) = 4/32;\ P(\texttt{d}) = 8/32;\ P(\texttt{e}) = 8/32.$$

A entropia para M_1 é calculada diretamente pela expressão de Shannon (3.2):

$$H(P) = -\sum_{\sigma \in \Sigma} P_\sigma \log_2 P_\sigma.$$

O cálculo da entropia para S a partir da distribuição de probabilidades fornecida pelo modelo M_1 resultou em $H(M_1) \approx 2,28$ bits/símbolo.

Seja M_2 o modelo aplicado pelo processo de consideração da existência de quatro estados vinculados à mensagem. No primeiro estado, somente são considerados os símbolos que ocorrem

nas posições $1, 5, 9, 13, \ldots$ de S. No segundo estado, somente são considerados os símbolos que ocorrem nas posições $2, 6, 10, 14, \ldots$ de S. No terceiro estado, somente são considerados os símbolos que ocorrem nas posições $3, 7, 11, 15, \ldots$ de S. No quarto estado, somente são considerados os símbolos que ocorrem nas posições $4, 8, 12, 16, \ldots$ de S. O tamanho da mensagem S é 32; assim cada estado ocorre exatamente 8 vezes. A probabilidade de cada estado ocorrer é 0,25. Será adotada a notação $P(e, \sigma)$ para denotar a probabilidade do símbolo σ no estado e. A distribuição das probabilidades para cada símbolo é:

Probabilidades de ocorrer cada símbolo em cada estado

Estado	a	b	c	d	e
1	$P(1, \mathtt{a}) = 1/8$	$P(1, \mathtt{b}) = 5/8$	$P(1, \mathtt{c}) = 0/8$	$P(1, \mathtt{d}) = 1/8$	$p(1, \mathtt{e}) = 1/8$
2	$P(2, \mathtt{a}) = 5/8$	$P(2, \mathtt{b}) = 0/8$	$P(2, \mathtt{c}) = 0/8$	$P(2, \mathtt{d}) = 3/8$	$P(2, \mathtt{e}) = 0/8$
3	$P(3, \mathtt{a}) = 0/8$	$P(3, \mathtt{b}) = 0/8$	$P(3, \mathtt{c}) = 0/8$	$P(3, \mathtt{d}) = 1/8$	$P(3, \mathtt{e}) = 7/8$
4	$P(4, \mathtt{a}) = 0/8$	$P(4, \mathtt{b}) = 1/8$	$P(4, \mathtt{c}) = 4/8$	$P(4, \mathtt{d}) = 3/8$	$P(4, \mathtt{e}) = 0/8$

A distribuição de probabilidades fornecida pelo modelo M_2 para S foi melhor que a obtida pelo modelo M_1, sendo menos uniforme. O cálculo da entropia para o modelo M_2 não é tão direto quanto o utilizado para o modelo M_1. É necessário calcular a entropia de cada estado $H(e)$ e então fazer o somatório das entropias de cada estado pela probabilidade do estado [5, 47], como visto na subseção entropia do processo de Markov (seção 3.4, equação (3.5)):

$$H(M_2) = \sum_{e=1}^{4} P(e)H(e)$$

sendo $P(e)$ a probabilidade de ocorrer o estado e em S e $H(e)$ a entropia do estado e, ou seja (ver a equação (3.4)),

$$H(e) = -\sum_{\sigma \in \Sigma} P(e, \sigma) \log_2 P(e, \sigma) \tag{4.1}$$

Desta forma, obtêm-se a entropia da distribuição de probabilidades fornecida pelo modelo M_2 para S calculando-se

$$H(M_2) = -\sum_{e=1}^{4} \sum_{\sigma \in \Sigma} P(e)P(e, \sigma) \log_2 P(e, \sigma)$$

sendo obtido o valor $H(M_2) \approx 1,11$ bits por símbolo.

A expressão generalizada para o cálculo da entropia para este tipo de modelo pode ser feita por:

$$H(M) = \sum_{e \in E} P(e)H(e) \tag{4.2}$$

e a partir de (4.1) e (4.2) obtém-se a expressão

$$H(M) = -\sum_{e \in E} \sum_{\sigma \in \Sigma} P(e)P(e, \sigma) \log_2 P(e, \sigma) \tag{4.3}$$

sendo M um modelo que gerou alguma distribuição de probabilidades e E o conjunto de todos os estados possíveis em alguma mensagem S qualquer.

Este exemplo deixou claro que a escolha no modelo interfere diretamente na estimativa da entropia da mensagem. A entropia da mensagem utilizando M_2 foi menos da metade da entropia da mesma mensagem utilizando M_1. Não faz sentido desvincular a estimativa da entropia do modelo adotado. Mas a escolha do modelo depende da mensagem na qual ele será utilizado. O conhecimento da mensagem possibilita a escolha de um modelo mais adequado e com isto se reduz significativamente sua estimativa da entropia.

A classificação de modelos, que será detalhada nas próximas seções, está intimamente ligada ao processo de compressão de uma mensagem, ilustrado na figura 4.1.

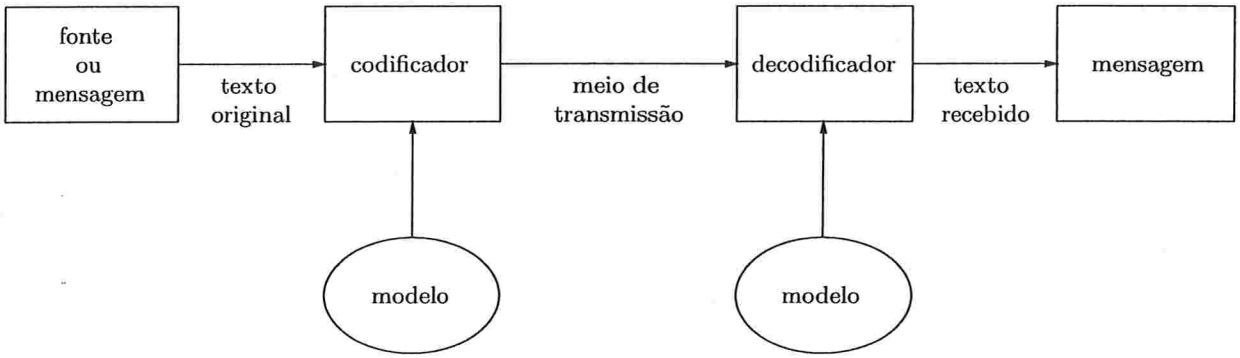


Figura 4.1: Esquema do processo de compressão.

4.1 Estáticos, adaptativos e semi-adaptativos

Modelos podem ser classificados em estáticos, semi-adaptativos ou adaptativos, dependendo da forma de cálculo utilizada para obtenção das distribuições de probabilidades.

Os modelos *estáticos* são aqueles previamente conhecidos pelo transmissor e pelo receptor. Deste modo tanto a codificação quanto a decodificação independem da mensagem. Esta forma de definir o modelo tem por principal vantagem a inexistência de perda de tempo no processo de manutenção do modelo. Um exemplo clássico do modelo estático é o código Morse.

No modelo *semi-adaptativo*, o transmissor lê toda a mensagem e cria um livro código com as palavras mais utilizadas, transmitindo para o receptor o livro código e em seguida a mensagem. Neste caso há um custo para a manutenção do modelo (livro código) mas há uma melhor taxa de compressão, uma vez que os códigos do livro refletem a mensagem para a qual foram criados.

No modelo *adaptativo*, como nos modelos estáticos, há um modelo inicial que existe tanto no

transmissor quanto no receptor e, à medida que o código é transmitido, ambos alternam da mesma forma o modelo, havendo um custo para sua manutenção. Normalmente existe a necessidade de uma estrutura complexa para atender de forma eficiente às alterações dinâmicas do modelo. *Envio* é a transmissão da mensagem do transmissor para o receptor. É comum na utilização dos modelos adaptativos o envio símbolo a símbolo, para que tanto o transmissor quanto o receptor possam atualizar suas distribuições de probabilidades. Somente as probabilidades associadas a uma classe de condição e as próprias classes de condição (seção 3.2) podem ser alteradas nos processos adaptativos.

Nos modelos adaptativos há o *problema da freqüência zero*, um custo adicional para representar os símbolos que ainda não apareceram. Eis um paradoxo clássico proposto por C. S. Pierce (1878): é dada uma urna com 1000 bolas de diversas cores, dentre elas as cores branca e preta. Sendo tirada ao acaso uma bola da urna, pergunta-se: "A bola é preta?" Há duas respostas possíveis: sim ou não. O método adaptativo induz a probabilidades de 1/2 para cada caso. Mas se a pergunta for: "A bola é preta, branca ou de outra cor?" Agora há três possíveis respostas associadas à probabilidade da bola ser preta, portanto esta é de 1/3. Isto contradiz a probabilidade de 1/2 da pergunta anterior. Para a mesma urna, com as mesmas bolas, nas mesmas condições [5]. O problema da freqüência zero somente é significativo quando se trabalha com ordens grandes (vide seção 4.2 sobre ordem do modelo) mas este problema pode ser resolvido com a utilização das técnicas de blending (detalhadas na seção 4.6).

Quando na mensagem ocorrerem diversas mudanças de contextos que inutilizem o que já foi 'aprendido' pelo modelo, a utilização de modelos adaptativos pode ser ruim. Seja considerada a mensagem

```
''aaaacccccgggggggttttttta''
```

e que nela seja utilizado um compressor adaptativo. A probabilidade do primeiro símbolo depende de como se faz a pergunta sobre ele (paradoxo de Pierce), ou seja, do modelo inicial adotado tanto no transmissor quanto no receptor. Quando o símbolo ainda não apareceu a princípio, sua probabilidade é zero ou próxima disto. Nesta mensagem é possível verificar que seu contexto muda toda vez que o modelo está se adaptando a uma certa situação, como se houvesse uma ruptura com o modelo anterior. Isto provoca baixas taxas de compressão.

Outra característica importante dos modelos adaptativos é: dado um conjunto de classes condicionais com k membros, um alfabeto de λ símbolos e uma mensagem de tamanho n, os modelos adaptativos possuem um limite para o tamanho da representação da mensagem em função de k, λ e n. Já os modelos não adaptativos podem expandir uma mensagem arbitrariamente [5].

Nem sempre é possível fazer alguma distribuição de probabilidades da mensagem antes de comprimi-la e transmiti-la para o receptor. Neste caso, a melhor opção é pela utilização de códigos adaptativos. O risco do trabalho com modelos estáticos é grande, ao passo que nos

modelos adaptativos as perdas, se ocorrerem, são limitadas. O importante aí é que tanto o transmissor quanto o receptor trabalhem com o mesmo modelo inicial, garantindo assim que a mensagem será corretamente decodificada.

Quando é impossível a criação do livro código das mensagens, a melhor opção é a utilização de métodos adaptativos. Estes, por sua vez, provocam um custo computacional extra. Por outro lado, quando há a possibilidade da criação do livro código, este deverá ser transmitido para o receptor, o que também é um custo. Dependendo do modelo, da ordem utilizada e do tamanho da mensagem, este custo pode ser até mais alto que o custo computacional da utilização dos modelos adaptativos.

4.2 Contextos finitos e ordem do modelo

No *modelo de contextos finitos*, a probabilidade do próximo símbolo é determinada pela seqüência de alguns dos símbolos antecessores imediatos. Este modelo também é chamado de modelo de Markov [5], já que os *contextos* podem ser entendidos como os estados do processo de Markov (seção 3.2) [47]. A *ordem do modelo* é o número de símbolos antecessores que são considerados para determinar a probabilidade do próximo símbolo. Por exemplo, na ordem 3 são considerados os 3 antecessores imediatos. Há duas ordens que fogem a esta regra: uma é a ordem -1, utilizada quando não há informações sobre a mensagem para a qual é feita uma distribuição equiprovável entre os símbolos do alfabeto. A outra é a ordem zero; a distribuição de probabilidades é obtida pela contagem das ocorrências de cada símbolo na mensagem. Para estas duas ordens em específico, não são levados em conta símbolos antecessores. O número de contextos para um modelo varia proporcionalmente ao tamanho do alfabeto e exponencialmente com a ordem do modelo.

Quando é utilizado algum modelo de ordem i, os primeiros i símbolos da mensagem não podem ser codificados por este, já que estes símbolos não possuem i antecessores. Neste caso, tanto o codificador quanto o decodificador devem possuir um modelo inicial para a representação dos primeiros i símbolos da mensagem. Para fins de compressão de dados, não é necessário que a ordem do processo de Markov seja utilizada no modelo. Uma seqüência pode seguir a centésima ordem de Markov e ser utilizado um modelo de contextos finitos de ordem 5. Desta forma, haverá alguma perda na previsibilidade do processo, mas haverá um ganho significativo na compressão da mensagem. No modelo de ordem 5 há λ^5 contextos e bem menos espaço é gasto para sua representação do que no modelo de ordem 100 com λ^{100} contextos. Na ordem 100 há muitas ocorrências do problema da freqüência zero, pois nem todos os contextos ocorrem. Por exemplo, na língua portuguesa, utilizando-se um modelo de contextos finitos de ordem 3, o contexto `ncp` não existe, já o contexto `nca` ocorre, por exemplo, na palavra `banca`. Quanto maior a ordem, maior será o número de contextos não utilizados. A escolha da ordem do modelo influencia na

distribuição das probabilidades e portanto no cálculo da entropia da mensagem.

Seja o alfabeto $\Sigma = \{\mathtt{a},\mathtt{t}\}$, $\lambda = 2$, $P(\mathtt{x})$ a probabilidade de algum símbolo, e a mensagem

$$S = ``\mathtt{attaattaaaaaatattaatttatttttaatt}''.$$

No modelo de ordem -1, como $\lambda = 2$ e

$$P(\mathtt{a}) = P(\mathtt{t}) = 0{,}5$$

O cálculo da entropia feito pela expressão de Shannon (3.2): $H(S) = -\sum_{\sigma\in\Sigma}(P_\sigma \log_2 P_\sigma)$ resulta em $H(S) = 1$ bit por símbolo.

No modelo de ordem 0

$$P(\mathtt{a}) = 15/32 \quad \text{e} \quad P(\mathtt{t}) = 17/32$$

O cálculo da entropia também é feito pela expressão de Shannon (3.2):
$H(S) = -\sum_{\sigma\in\Sigma}(P_\sigma \log_2 P_\sigma)$ resulta em $H(S) = 0{,}997$ bit por símbolo.

A partir da ordem 1 começam as considerações sobre os símbolos antecessores. O cálculo da distribuição das probabilidades nos modelos de ordem superior a 0 requer uma consideração sobre os antecessores imediatos. Será adotada a notação $P(\sigma|s)$ para indicar a probabilidade de ocorrer o símbolo σ no contexto s. Para o exemplo acima, na ordem 1, há dois contextos possíveis: a e t.

$$P(\mathtt{a}|\mathtt{a}) = 8/15;\ P(\mathtt{t}|\mathtt{a}) = 7/15;\ P(\mathtt{a}|\mathtt{t}) = 6/17;\ p(\mathtt{t}|\mathtt{t}) = 10/17.$$

Há duas possibilidades para tornar unitárias as probabilidades após a ocorrência do evento t. A primeira é diminuir 1 do número de possíveis cadeias antecessoras. A segunda é incluir no alfabeto um símbolo especial para o fim da mensagem. No exemplo anterior, a diferença ocorre devido ao término da mensagem com o símbolo t, não havendo sucessor para esta ocorrência. Adotando a primeira solução tem-se a seguinte distribuição de probabilidades, para o modelo de ordem 1:

$$P(\mathtt{a}|\mathtt{a}) = 8/15;\ P(\mathtt{t}|\mathtt{a}) = 7/15;\ P(\mathtt{a}|\mathtt{t}) = 6/16;\ P(\mathtt{t}|\mathtt{t}) = 10/16.$$

O cálculo da entropia de um modelo de contextos finitos de ordem k depende do cálculo prévio da entropia de cada contexto s, de tamanho k, que é dada por (como na equação (3.4)):

$$H(s) = -\sum_{\sigma\in\Sigma} P(\sigma|s)\log_2 P(\sigma|s). \tag{4.4}$$

Após o cálculo da entropia de cada contexto, a entropia do modelo é obtida por (como na equação (3.5)):

$$H(S) = \sum_{s\in S} P(s)H(s) \tag{4.5}$$

sendo $P(s)$ a probabilidade de ocorrer o contexto s de tamanho k na mensagem S. Calculando a entropia de S, a partir da distribuição de probabilidade para a ordem 1 obtida acima, pelas expressões (4.4) e (4.5): $H(S) = 0{,}944$ bit por símbolo.

No modelo de ordem 2 as distribuições de probabilidade são:

$$P(\mathtt{a}|\mathtt{aa}) = 4/8,\ P(\mathtt{t}|\mathtt{aa}) = 4/8;\ P(\mathtt{a}|\mathtt{at}) = 1/7;\ P(\mathtt{t}|\mathtt{at}) = 6/7;$$
$$P(\mathtt{a}|\mathtt{ta}) = 4/6;\ P(\mathtt{t}|\mathtt{ta}) = 2/6;\ P(\mathtt{a}|\mathtt{tt}) = 5/9;\ P(\mathtt{t}|\mathtt{tt}) = 4/9.$$

O cálculo da entropia para a distribuição de probabilidades feito pelas expressões (4.4) e (4.5) resultou em $H(S) = 0{,}886$ bit por símbolo. A distribuição de probabilidades para a ordem 3 possui 16 ocorrências e na ordem 4 há 32. Embora haja apenas dois símbolos no alfabeto deste exemplo, o tamanho da mensagem (32 símbolos) ficará menor que o tamanho da distribuição das probabilidades, se a ordem do modelo for maior que 4. A distribuição das probabilidades para a ordem 2 é menos uniforme que para a ordem 1, o que gera um menor valor para a estimativa da entropia da mensagem.

É crucial no modelo de contextos finitos o fato de que a distribuição de probabilidades associada ao modelo é baseada em um certo número (ordem) de símbolos antecessores ao que se deseja prever. Mas para que a mensagem possa ser decodificada, é necessário algum custo para a representação da distribuição das probabilidades e, dependendo da ordem escolhida, isto pode comprometer consideravelmente a taxa de compressão (seção 3.3). O modelo de contextos finitos pode ser interpretado como um processo de Markov (seção 3.2).

4.3 Estados finitos e ergódico

Em algumas situações, o conhecimento prévio de um certo número de antecessores não diminui a incerteza quanto ao próximo símbolo. Nestes casos, o conhecimento adicional de algumas condições pode ajudar na diminuição da incerteza. O modelo M_2 no início deste capítulo é um exemplo onde o conhecimento adicional ajudou na previsibilidade da mensagem. Em geral quando ocorrem "estados" há esta ajuda, pois para cada estado existe uma distribuição de probabilidades.

Uma mensagem pode ser decomposta em um número finito de estados se cada símbolo da mensagem pertencer a exatamente um estado. Um *estado* é uma situação específica para a qual há uma distribuição de probabilidades para os símbolos do alfabeto e outra distribuição de probabilidades para os estados possíveis da mensagem. Mais precisamente, um modelo de estados finitos é composto por quatro objetos: um conjunto finito de E estados, um alfabeto Σ, um estado inicial k em E, um subconjunto K_f de E de estados finais e uma função de transição δ de $E \times \Sigma$ em E.

Considere um exemplo de modelo de estados finitos a posição do nucleotídeo no códon, em uma seqüência de códons. Há 3 estados possíveis, $E = \{1, 2, 3\}$, $\Sigma = \{$ a, c, g, t $\}$. O estado inicial é 1 e o final 3. A função de transição é $\delta(1, \sigma) = 2$, $\delta(2, \sigma) = 3$, $\delta(3, \sigma) = 1$, σ em Σ. Este modelo está ilustrado na figura 4.2.

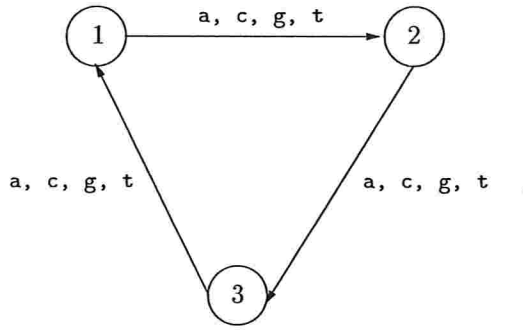


Figura 4.2: Modelo de estados finitos representando códons.

A primeira análise com relação aos estados será feita em função das probabilidades de transição de um estado para outro. Seja $P(i \rightarrow j)$ a probabilidade de transição do estado i para o estado j. Como a informação no DNA é linear, a probabilidade de cada estado é 1/3 e a distribuição de probabilidades entre estados é:

$$P(1 \rightarrow 1) = 0;\ P(1 \rightarrow 2) = 1;\ P(1 \rightarrow 3) = 0$$
$$P(2 \rightarrow 1) = 0;\ P(2 \rightarrow 2) = 0;\ P(2 \rightarrow 3) = 1$$
$$P(3 \rightarrow 1) = 1;\ P(3 \rightarrow 2) = 0;\ P(3 \rightarrow 3) = 0$$

Outra análise está na distribuição das probabilidades dentro de cada estado. Não será produtivo quebrar uma mensagem em estados se não houver diminuição da incerteza ao se fazer isto.

O cálculo da entropia do modelo de estados finitos se faz através da fórmula (4.3), de maneira similar à utilizada para calcular a entropia do modelo M_2, do início deste capítulo.

O estado normalmente está associado a uma condição "real", o *modelo físico*, que é o modelo gerador da mensagem. Quando este modelo é conhecido, ele costuma ser utilizado como o modelo adotado para gerar a distribuição de probabilidades da mensagem. Se para uma porção contígua do DNA é válido um modelo com dois estados $E = \{1, 2\}$, com as seguintes características: $\Sigma =$ $\{$ a,c,g,t $\}$, o estado inicial é 1 e o final é 2 e δ(1,a= 1) δ(1,c= 1) δ(1,g= 2) δ(1,t= 2) δ(2,a= 2) δ(2,c= 2) δ(2,g= 1) δ(2,t= 1), como mostrado na figura 4.3.

A mensagem abaixo, a partir da qual foi criado este modelo de estados finitos, possui as

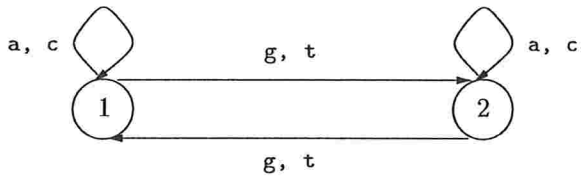


Figura 4.3: Exemplo de modelo com dois estados.

seguintes probabilidades para os estados: $P(1) = 5/8$ e $P(2) = 3/8$.

''aactagcactgggctatccctattgccagcag''

O que está sublinhado na mensagem acima está no estado 1, o não sublinhado no estado 2. As distribuições de probabilidade para este modelo de estados finitos de ordem zero são:

$$P(1,\mathtt{a}) = 3/20 \; P(1,\mathtt{c}) = 10/20 \; P(1,\mathtt{g}) = 2/20 \; P(1,\mathtt{t}) = 5/20$$
$$P(2,\mathtt{a}) = 4/12 \; P(2,\mathtt{c}) = 1/12 \; P(2,\mathtt{g}) = 5/12 \; P(2,\mathtt{t}) = 2/12$$

e sua entropia, calculada a partir da expressão (4.3), é $H(M) = 1{,}758$. As distribuições de probabilidade para o modelo de contextos finitos de ordem zero, nesta mesma mensagem, são:

$$P(\mathtt{a}) = 7/32 \; P(\mathtt{c}) = 11/32 \; P(\mathtt{g}) = 7/32 \; P(\mathtt{t}) = 7/32$$

tendo sua entropia calculada, pela equação de Shannon (3.2) em 1,968 bits por símbolo. O conhecimento da existência de algum modelo de estados finitos gerando a mensagem, diminuiu sua entropia (incerteza) e aumentou sua previsibilidade.

Os modelos de contextos podem ser entendidos como modelos de estados finitos onde cada contexto (seqüência de um certo número de símbolos) pode representar um estado. Nem todo modelo de estados finitos é um modelo de contextos finitos. No exemplo acima, as subcadeias de tamanho três `ggg` e `ttg` aparecem na mensagem, mas não formam um contexto, pois cada `g` e cada `t` mudam o estado da mensagem.

Modelo ergódico

Um modelo de estados finitos é *ergódico* (ver seção 3.2) se a probabilidade for maior que zero para chegar a algum estado j, a partir de algum estado i, para quaisquer estados i e j. Em outras palavras, para qualquer estado inicial da mensagem é possível a chegada a cada um dos demais estados do modelo, caso a mensagem seja suficientemente longa.

Para os exemplos seguintes o símbolo "|" representa "ou" e o símbolo "$\rightarrow$" representa que a probabilidade de sair do estado a esquerda do símbolo para qualquer um dos estados a direita

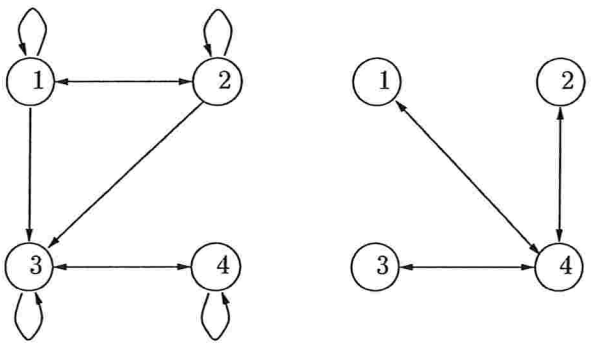
do símbolo é não nula, sendo nula para os estados não representados à direita do "$\rightarrow$". Dado um modelo com quatro estados possíveis $E = \{1, 2, 3, 4\}$, o seguinte modelo não é ergódico, pois uma seqüência começada nos estados 3 ou 4 jamais atingiria os estados 1 ou 2 por mais que ela crescesse.

$$\begin{aligned} 1 &\rightarrow 2 \mid 3 \mid 1 \\ 2 &\rightarrow 2 \mid 3 \mid 1 \\ 3 &\rightarrow 3 \mid 4 \\ 4 &\rightarrow 3 \mid 4 \end{aligned}$$

Já o próximo modelo é ergódico, pois se uma dada seqüência for suficientemente grande esta poderá representar todos os estados do modelo $1, 2, 3, 4$, qualquer que seja o estado inicial ou decisão tomada no meio da seqüência.

$$\begin{aligned} 1 &\rightarrow 4 \\ 2 &\rightarrow 4 \\ 3 &\rightarrow 4 \\ 4 &\rightarrow 1 \mid 2 \mid 3 \end{aligned}$$

Os exemplos acima encontram-se ilustrados na Figura 4.4.



Modelo ergódico Modelo não-ergódico

Figura 4.4: Exemplos de modelos ergódicos e não-ergódicos.

Todos os modelos de estados finitos vinculados à linguagem natural são ergódicos. O modelo ser ergódico, ou não, influencia na estimativa da entropia. Se o modelo é ergódico, sua distribuição de probabilidades tende a independer do estado inicial, se a mensagem for suficientemente longa. No primeiro exemplo acima para os estados $1, 2, 3, 4$ suponha que as probabilidades dos

estados sejam: $P(1) = 0{,}6$; $P(2) = 0{,}3$; $P(3) = 0{,}08$ e $P(4) = 0{,}02$; uma mensagem qualquer iniciada em 1 ou 2 teria uma entropia muito diferente de uma mensagem iniciada em 3 ou 4, uma vez que não é possível para mensagens iniciadas por 3 ou 4 atingir os estados 1 ou 2, apesar de $P(1) + P(2) = 0{,}9$. Já no segundo exemplo $P(4) = 0{,}5$.

Para modelos ergódicos a entropia do modelo como um todo é a soma das entropias individuais de cada estado, levando-se em conta a probabilidade do estado (equação (4.2)).

4.4 Gramáticos

São modelos que definem bem linguagens de computação como Cobol, Pascal ou qualquer linguagem de sintaxe rígida e bem definida. Estes modelos não são bons para representar a linguagem natural, pois embora exista uma sintaxe, ela é variável e cheia de exceções como frases sem verbos (Fogo!), frases sem sujeito (Faz frio.), frases incompletas (Eu gostaria, mas...), verbos irregulares como o verbo ir, frases intercaladas (João, que é o homem forte da casa, tremeu de medo.), frases que completam frases (Ele queria que seu dinheiro desse para viajar.), etc.

Uma *gramática* é composta por quatro ingredientes: um alfabeto V, um subconjunto Σ de V de *símbolos terminais*, um *símbolo inicial* S em $V - \Sigma$ e um conjunto R de *regras* que é algum subconjunto de $(V^*(V - \Sigma)V^*) \times V^*$. Os elementos de $V - \Sigma$ são chamados de *símbolos não terminais*.

No artigo de Lanctot, Li and Yang [28] é feita uma transformação gramática na seqüência do DNA antes da utilização de um compressor aritmético (seção 5.4) de ordem zero para comprimir a gramática e a mensagem resultante da transformação gramática (capítulo 9).

4.5 Modelagem da linguagem natural

A modelagem da linguagem natural é bastante utilizada por diversos compressores e é muito comum no ambiente de informática. Embora não haja relações diretas entre a linguagem natural e o DNA, o desenvolvimento das técnicas de compressão passaram pela compressão da linguagem natural e as ferramentas vindas da linguagem natural são freqüentemente utilizadas nas técnicas de compressão para o DNA, porém sem a eficiência que marca a compressão da forma mais habitual de comunicação do ser humano.

No artigo de Ebeling e Frömmel [17] é detectado um aspecto comum à linguagem natural e cadeias primárias de DNA: as correlações de longa distância. Mas o aspecto mais relevante da modelagem da linguagem natural está na diversidade de modelos e técnicas de compressão criados com o fim de comunicação. Um exemplo básico é a linguagem Braille, para deficientes

visuais, que mistura letras, palavras e abreviações em seu alfabeto [5].

A linguagem natural é composta de orações, que são formadas de palavras, espaços em branco e sinais de pontuação. As palavras são escritas com letras e sinais como apóstrofe, hífens e acentuação. Para se comprimir textos em linguagem natural é necessário definir qual o tipo e a ordem do modelo a ser adotado. Os conceitos de letra e palavra são adaptados para atender à modelagem dos textos em linguagem natural. As *letras* são quaisquer caracteres válidos. *Palavra* é uma seqüência finita de símbolos que não contenham o espaço em branco e suas variações (tabulação, parágrafo, quebra de linha, etc). Porém o conceito de palavra não é unânime, podendo possuir outras definições em outras publicações. Os *digramas* são um conjunto de duas letras ou palavras dependendo do contexto. Os *trigramas* e *tetragramas* são respectivamente um conjunto de três ou quatro letras ou palavras dependendo do contexto. Em se tratando de linguagem natural, nem todos os digramas, trigramas e tetragramas ocorrem. Na língua portuguesa o digrama `çç` não ocorre nem o trigrama `nhm` e o tetragrama `arrc`. Estudos feitos para a língua inglesa, num alfabeto de 94 letras, mostram que ocorrem apenas 39% dos digramas possíveis; 3,6% dos trigramas possíveis e 0,2% dos tetragramas possíveis. Na criação de textos aleatórios, a partir de estatísticas de digramas, trigramas, ..., 12-gramas há uma aproximação progressiva do inglês, porém sem nenhum sentido.

George Zipf publicou em 1949 um livro no qual o autor determina que as palavras mais utilizadas na língua inglesa seguem o princípio do menor esforço. Ele achou uma hipérbole para a freqüência da utilização das palavras. Foi um trabalho não muito científico e sem grande peso.

G.A. Miller se utilizou de um macaco para digitar palavras aleatoriamente, seguindo somente duas restrições: a) a barra de espaço teria a probabilidade P e as demais letras $1-P$; b) o macaco jamais digitaria duas vezes a barra de espaço na mesma linha. O macaco gerou uma saída cuja estatística ficou bem próxima à de Zipf. O modelo de Zipf não é preciso para uma distribuição de letras. Embora seu trabalho não seja muito preciso, ele foi o primeiro a utilizar a distribuição hiperbólica.

4.6 Blending

Blending[1] é uma técnica de modelagem que permite que diferentes ordens sejam utilizadas simultaneamente para determinar a probabilidade do próximo símbolo. Nesta técnica, essencialmente adaptativa, cada ordem ganha um peso e a ordem -1 é usada para permitir a utilização dos símbolos do alfabeto que ainda não ocorreram. A soma dos pesos das ordens deve ser 1. A técnica de blending normalmente é empregada para fornecer as distribuições de probabilidades para algum algoritmo, como o algoritmo de compressão aritmética. Duas características impor-

[1] Palavra inglesa que significa mistura.

tantes do blending são: a existência de um símbolo de escape e um limite máximo para a ordem a ser adotada. É utilizada normalmente a letra m para indicar este limite máximo.

Por exemplo, se o contexto `ata` já ocorreu 10 vezes, sendo sucedido por `a` duas vezes, por `c` uma vez, por `g` cinco vezes e por `t` duas vezes, as probabilidades são: $P(\texttt{a}) = 0{,}2$, $P(\texttt{c}) = 0{,}1$, $P(\texttt{g}) = 0{,}5$ e $P(\texttt{t}) = 0{,}2$. Se o sucessor é `t` é transmitido um código com tamanho $-\log_2(0{,}2)(\approx 2{,}3)$ bits. Porém se o `t` nunca houvesse ocorrido no contexto `ata` a técnica blending de tenta utilizar o contexto reduzido `ta`, informando ao decodificador através da transmissão do caractere de escape. O processo se repete até a ordem -1, para a qual não há uma probabilidade associada ao caractere de escape.

Há várias formas de se calcular a probabilidade do escape, a qual influencia nos pesos de cada ordem. A grande dificuldade da estrutura da técnica de blending pura está na forma complexa para se manter dinamicamente as probabilidades e os pesos. Uma solução usual está na adoção da simplificação, chamada exclusão, que considera a probabilidade de cada símbolo ocorrer somente para o contexto de ordem mais alta para a qual ele esteja previsto. Vários algoritmos foram criados utilizando blending, mas será dado destaque somente ao *PPM*, sigla em inglês para previsões com conferência parcial (*prediction with partial match*). Há diversas variações para o PPM, sendo a principal diferença entre elas a forma de se calcular o caractere de escape.

PPMA

Sayood [47] utiliza o PPMA em conjunto com a compressão aritmética para exemplificar as tabelas de controle necessárias para o algoritmo. No PPMA a técnica utilizada para calcular os escapes consiste em incluir um caractere de escape que é transmitido (utilizando a compressão aritmética, por exemplo) quando, naquela ordem, o próximo símbolo ainda não ocorreu. O símbolo de escape ocorre para todas as ordens, exceto a ordem -1. Este modelo é muito prático quando há pouca necessidade de se recorrer ao símbolo de escape, pois deixa relativamente pouca probabilidade para sua utilização.

PPMB

O PPMB utiliza para o cálculo do escape o seguinte fato: a primeira vez que um símbolo ocorre ele não é incluído nas probabilidades e esta ocorrência é adicionada ao contador de escapes. Desta forma, o contador de escapes terá exatamente o número de símbolos que ocorreu e cada símbolo terá seu contador reduzido na unidade em cada ordem. A vantagem deste método é que os casos anômalos, que ocorrem somente uma vez, não entram na distribuição de probabilidades, deixando-a direcionada aos eventos que ocorreram mais de uma vez.

PPMC

O PPMC assim como o PPMB utiliza para o contador de escape o número de símbolos que apareceram para a ordem, mas sem subtrair em um o contador de cada símbolo.

Note-se que, qualquer que seja o algoritmo utilizado, a distribuição de probabilidades sofrerá a exclusão para que cada símbolo tenha somente a probabilidade da mais alta ordem na qual ele apareça. Este é o princípio do blending que serve para indicar as probabilidades para algum algoritmo.

Os modelos ditam as distribuições de probabilidades das mensagens, definindo a entropia e facilitando (ou dificultando) suas compressibilidades. Alguns modelos tornam algumas mensagens mais previsíveis que outros, mas cada modelo tem sua utilização, de acordo com a mensagem alvo. Vários modelos possuem aspectos históricos relevantes. Serão vistos no capítulo 5 os principais algoritmos que utilizam os modelos e outros tipos de compressores.

Capítulo 5

Técnicas de Compressão de Dados

Há várias formas de se comprimir dados, cada qual partindo de um princípio básico que seu autor acha mais eficiente. Cada método possui suas vantagens, desvantagens e campo de aplicação. É raro ser possível afirmar que o princípio da técnica "A" é melhor ou pior que o princípio da técnica "B". Um exemplo está no código Morse, inventado em 1835 por Samuel Morse. O princípio adotado por Morse para definir seu código foi: caracteres mais utilizados devem ser mais rapidamente transmitidos que os menos utilizados. Foi criado um alfabeto binário composto por ponto e traço (três vezes mais lento que o ponto). Desta forma o `e` é representado somente por um ponto (pulso mais rápido) e o `j` por um ponto e três traços. Os algarismos utilizam cinco pulsos (traços ou pontos). Entre cada símbolo de um mesmo caractere há um espaço (silêncio) do tamanho (tempo) de um ponto. A vantagem do código Morse está na eficiência da transmissão de mensagens. A desvantagem está na pouca versatilidade para lidar com expressões matemáticas. Outra característica deste código é que um símbolo pode ser prefixo de outro (como o `e` é prefixo do `j`). Após o término da transmissão de cada letra é feita uma pausa equivalente a um traço para indicar o final desta; ao término de cada palavra é colocada uma pausa de sete pontos. Morse utilizou um modelo de contextos finitos, estático e de ordem zero. O escopo do código Morse ficou na transmissão de mensagens no telégrafo e uso militar, mas é história e não possui mais utilização. Os conhecimentos necessários para a compreensão deste capítulo estão descritos nos capítulos 3 e 4.

5.1 Código de Shannon-Fano

O código de Shannon-Fano é parecido com o do código Morse pois compartilha o mesmo princípio. Adaptado à tecnologia computacional, representa cada mensagem sobre um alfabeto $\{0, 1\}$. O algoritmo básico deste código recebe um alfabeto Σ e uma distribuição de probabilidades sobre os símbolos deste. Devolve uma codificação para cada símbolo sobre o alfabeto $\{0, 1\}$. O algoritmo é recursivo e no início de cada iteração são aplicados os seguintes passos:

- os símbolos da lista são ordenados de acordo com suas probabilidades;
- a lista é particionada em duas sublistas cuja soma das probabilidades é aproximadamente a mesma;
- a codificação dos símbolos da primeira sublista é iniciada com 0 e a da segunda com 1;
- o algoritmo continua recursivamente até que cada lista possua apenas 1 símbolo.

No final da última iteração, cada símbolo será representado por uma seqüência de zeros e uns. É possível que símbolos equiprováveis possuam tamanhos distintos quando codificados por este método. Neste tipo de algoritmo existe a *redundância*, códigos que não representem nenhum símbolo. Aplicando o código de Shannon-Fano na mensagem

$$\texttt{``aacaagccaattcaggatag''}$$

as probabilidades de cada símbolo são:

$$P(\texttt{a}) = 9/20,\ P(\texttt{c}) = 4/20,\ P(\texttt{g}) = 4/20,\ P(\texttt{t}) = 3/20$$

O algoritmo recebe as probabilidades e na primeira iteração separa o a (9/20) dos demais símbolos (11/20). Na próxima iteração separa o c (4/20) dos símbolos restantes(7/20) e na última iteração são representados os dois últimos símbolos g (4/20) e t (3/20). Assim o algoritmo devolverá:

$$\texttt{a} = \{0\},\ \texttt{c} = \{10\},\ \texttt{g} = \{110\},\ \texttt{t} = \{111\}$$

ficando a mensagem representada por

$$00100011010100011111110011011001110110$$

com a utilização de 38 bits.

Este código permite a representação de uma mensagem com um número menor de bits, mas tem o custo da administração de redundâncias. A redundância é inevitável; caso todos os códigos possíveis fossem utilizados haveria casos nos quais uma seqüência de bits possuiria mais de um significado. A compressão se tornaria irreversível[1] ou *ambígua*[2]. É fundamental que a representação de um símbolo não seja prefixo da representação de outro. O código de Shannon-Fano se baseia no modelo não adaptativo e cada símbolo é representado por um número fixo de bits; este pode variar de um símbolo para outro.

[1] Forma de compressão que não permite o retorno à mensagem original.

[2] Mais de uma forma de decodificação da mensagem; dentre elas a forma original que foi utilizada na codificação.

5.2 Código de Huffman

O código de Huffman apresenta uma reunião de dois princípios. O primeiro, utilizado pelo código de Shannon-Fano, visa a representação dos símbolos mais prováveis com menor número de bits. O segundo torna iguais o número de bits na representação dos dois símbolos menos prováveis. Possui as mesmas vantagens e desvantagens do código de Shannon-Fano, mas com algumas diferenças impostas pela segunda condição. O algoritmo do código de Huffman recebe um alfabeto Σ, uma distribuição de probabilidades sobre os símbolos do alfabeto, classificada em ordem crescente de probabilidades. Devolve uma codificação para cada símbolo sobre o alfabeto $\{0, 1\}$. O algoritmo é recursivo e no início de cada iteração são aplicados os seguintes passos:

- caso haja somente um símbolo na lista, termina a recursão e o algoritmo devolve a representação de cada símbolo, senão:
- os dois símbolos menos prováveis são representados; um deles como sendo final 0 e o outro como sendo final 1;
- é feita a reunião destes símbolos como se fossem um e somadas suas probabilidades;
- a lista é reordenada em ordem crescente de probabilidades;

Aplicando o código de Huffman na mensagem

$$\texttt{``aacaagccaattcaggatag''}$$

as probabilidades de cada símbolo são:

$$P(\texttt{a}) = 9/20,\ P(\texttt{c}) = 4/20,\ P(\texttt{g}) = 4/20,\ P(\texttt{t}) = 3/20$$

Na primeira iteração, o algoritmo recebe $P(\texttt{t}) = 3/20, P(\texttt{g}) = 4/20, P(\texttt{c}) = 4/20, P(\texttt{a}) = 9/20$. Como o número de símbolos é maior que um, o `t` é representado como final 1 e o `g` como final 0. O novo símbolo `gt` possui $P(\texttt{gt}) = 7/20$ e a lista é ordenada ficando $P(\texttt{c}) = 4/20, P(\texttt{gt}) = 7/20, P(\texttt{a}) = 9/20$ para o início da segunda iteração. Como o número de símbolos é maior que um, o `c` é representado como final 1 e o `gt` como sendo final 0. O novo símbolo `cgt` possui a probabilidade $P(\texttt{cgt}) = 11/20$. A nova lista é ordenada e fica $P(\texttt{a}) = 9/20, P(\texttt{cgt}) = 11/20$ para a terceira iteração. Como há mais de um símbolo, o `a` fica representado como final 1 e o `cgt` como final 0. O novo símbolo `acgt` possui probabilidade 1, e na quarta e última iteração o algoritmo devolve

$$\texttt{a} = \{1\},\ \texttt{c} = \{01\},\ \texttt{g} = \{000\},\ \texttt{t} = \{001\}$$

a mensagem codificada é representada por

$$110111000010111001001011000000010011000$$

sendo utilizados 38 bits. O `t` é representado por 001 pois o 1 é a representação do sufixo do `t` sozinho quando este apareceu como símbolo menos provável. Um 0 é a representação do sufixo do `gt` quando este esteve entre os símbolos menos prováveis e o outro 0 é a representação do sufixo do `cgt`. O conceito de final 0 e final 1 do primeiro passo do algoritmo significa que a representação do símbolo em questão termina com o bit 0 ou o bit 1, conforme foi atribuído pelo algoritmo.

Na média, o código de Huffman é mais eficiente (menor redundância) que o código de Shannon-Fano.

Os dois códigos anteriores são baseados em princípios da representação de cada símbolo por um número fixo de bits e na suposição de que é eficiente representar os símbolos/mensagens mais freqüentes com um número menor de bits. Mas os compressores devem ser adaptados para a realidade da mensagem que pretendem comprimir. Quando esta realidade é conhecida, ela pode melhorar as taxas de compressão. Como mostrado na próxima seção.

5.3 Um caso incomum

Seja uma mensagem a ser representada que possua 90.000.000 de números inteiros menores que 1.000.000.000. Neste caso o número de palavras é muito grande. Seja a hipótese de que estes números tenham a seguinte propriedade: em 98% dos casos o módulo da diferença de um número para seu sucessor é menor que 20. Dadas estas condições de contorno, o algoritmo abaixo atende bem à representação da mensagem.

- o primeiro número é codificado ou decodificado em código binário em seu tamanho máximo possível (30 bits).
- A partir do segundo número, a representação assume a seguinte regra para o codificador:
  - o próximo número é subtraído do último codificado;
  - se a diferença for maior ou igual a zero e menor que 32 ela é codificada nos próximos 6 bits;
  - se a diferença for menor que zero e maior que −31, o próximo bit é ligado e a diferença é representada nos 5 bits seguintes;
  - se não ocorrer nenhum dos dois casos anteriores, os 6 próximos bits são ligados e o número em binário é representado nos 30 bits seguintes.
- A partir do segundo número, a representação assume a seguinte regra para o decodificador:
  - os 6 próximos bits são examinados:

- se o primeiro bit estiver desligado, os próximos 5 bits representam o número em base 2 que deverá ser somado ao número atual para a decodificação do próximo número;
- se os próximos 6 bits estiverem todos ligados, a representação do próximo número estará em código binário nos 30 bits seguintes;
- se não ocorrer nenhum dos dois casos anteriores (o primeiro bit está ligado e pelo menos um dos outros 5 bits estiver desligado) os próximos 5 bits seguintes ao bit inicial ligado indicam o número binário que deverá ser subtraído do número atual para a decodificação do próximo número.

Este algoritmo, criado sob a encomenda do escopo acima, pode representar um dado número com a utilização de 6, 36 ou 30 (primeiro número) bits, conforme a situação na qual ele esteja. Mas como em 98% dos casos a diferença é menor que 20, a representação do número utilizará apenas 6 bits.

Uma das vantagens deste algoritmo está no fato de permitir a criação de um índice simples (a cada mil números por exemplo) que permitiria um acesso relativamente direto ao número 7.235.415 da mensagem que está sendo comprimida. Uma das desvantagens da compressão está no fato de quase nunca é possível o acesso direto ao dado. Mas no quotidiano as mensagens quase nunca são tão bem comportadas para que este algoritmo seja recomendável para uso geral. Caso exista alguma regra na formação da seqüência do DNA, a determinação desta regra tornaria a cadeia bem mais previsível e diminuiria sua entropia. Porém, se há tal regra, esta ainda não foi descoberta.

5.4 Compressão aritmética

A compressão aritmética é um exemplo que mostra a diversidade de compressores existentes. O princípio no qual se baseia este algoritmo é completamente diferente dos utilizados nos anteriores, embora sua finalidade seja a mesma. Na compressão aritmética, cada mensagem é representada por um número no intervalo entre [0, 1). Quanto maior a mensagem, menor o intervalo necessário para representá-la e maior a quantidade de algarismos (bits) para codificar este intervalo. Ao contrário dos algoritmos anteriores, este não permite a criação de um índice para o acesso direto. Outra diferença está na necessidade de haver um símbolo para representar o final da mensagem, sem o qual seria impossível decodificá-la.

O algoritmo de compressão aritmética recebe um alfabeto Σ, incluído nele um símbolo especial de fim de mensagem, uma distribuição de probabilidades sobre Σ e uma mensagem $S \in \Sigma^*$ terminada pelo símbolo de final de mensagem. Devolve uma codificação para S que representa um número no intervalo $[0, 1)$. Seja $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_\lambda\}$ e $P(i)$ a probabilidade associada ao símbolo i $(1, 2, \ldots, \lambda)$. O algoritmo é iterativo e no início de cada iteração o algoritmo possui

um intervalo $[a, b)$, uma distribuição de probabilidades sobre Σ, um símbolo $\sigma \in \Sigma$ de S. Na primeira iteração $a = 0$, $b = 1$, σ é o primeiro símbolo de S. Em cada iteração o algoritmo efetua os seguintes passos:

- Uma vez codificado algum símbolo, é feito o ajuste do intervalo para a inclusão do novo símbolo. Supondo $\sigma = \sigma_i$ e $P(0) = 0$, $a' = a + (b - a) \times \sum_{j=0}^{i-1} P(j)$ e $b' = a + (b - a) \times \sum_{j=0}^{i} P(j)$. Caso o algoritmo esteja trabalhando de modo adaptativo, são feitos os devidos ajustes na distribuição de probabilidade.

- Caso 1: o símbolo codificado não é o símbolo de final de mensagem. σ' recebe o próximo símbolo de S e é feita nova iteração com $a = a'$, $b = b'$, $\sigma = \sigma'$ e a distribuição de probabilidades ajustada, caso o algoritmo esteja trabalhando de modo adaptativo.

- Caso 2: o símbolo codificado é o símbolo de final de mensagem. Um valor dentro do intervalo final é escolhido convenientemente, o algoritmo devolve o valor codificado e pára.

O decodificador trabalha da mesma forma, recebe a distribuição de probabilidades idêntica à utilizada pelo codificador e, de acordo com a probabilidade da mensagem recebida, vai identificando cada símbolo enviado, até chegar ao símbolo de final de mensagem. Este algoritmo pode ser implementado para modelos estáticos, adaptativos ou para modelos semi-adaptativos.

Uma primeira consideração é sobre a precisão da máquina. Uma mensagem muito grande exigiria um número considerável de bits para representá-la. Desta forma seria inviável a implantação da compressão aritmética. Soluções independentes foram achadas quase simultaneamente por Rubin, Guazzo e Rissanen, no final da década de 70. A visão moderna da compressão aritmética utiliza números inteiros e leva em conta tanto a transmissão e recepção incremental quanto problemas como *overflow* e *underflow* e eficiência do algoritmo.

O trabalho com números inteiros consiste no trato das probabilidades através da utilização de intervalos de números inteiros. Por exemplo é possível o trabalho com números entre 0 (zero) 10.000, onde 0,072 (7,2%) é representado pelo número 720.

A transmissão e recepção incremental foi implementada para a versão binária do algoritmo e se baseia neste princípio: quando o bit de mais alta ordem dos números que representam os extremos do intervalo $[a, b)$ são os mesmos, estes não interferem mais no processo, podendo ser transmitidos. Toda vez que a e b estiverem na mesma metade de um número binário, os bits de mais alta ordem dos extremos do intervalo $[a, b]$ serão iguais. É feita a transmissão deste bit e os devidos ajustes no intervalo $[a, b)$.

A utilização da aritmética de ponto fixo exige que quanto maior a mensagem maior deverá ser a precisão do número a ser trabalhado para codificá-la e decodificá-la. Devem ser tomadas medidas para evitar o *overflow* quando se trabalha com um número maior do que o que se

consegue representar, e o *underflow* quando o intervalo $[a, b)$ não é grande o suficiente para representar um determinado código.

No processo da compressão aritmética, o codificador e o decodificador utilizam as mesmas consistências para evitar *underflow/overflow*. Uma forma de evitar o *underflow* é garantir que o tamanho do intervalo $(b - a)$ seja sempre maior que o tamanho do símbolo mais freqüente. Toda vez que o tamanho do intervalo ficar menor que o tamanho necessário para a representação deste símbolo um valor fixo é somado a b e subtraído de a. Com isto é garantida a condição suficiente para que não ocorra *underflow*. A implementação em binários garante que não haverá *overflow* pela utilização da freqüência máxima permitida e de outros conceitos que não cabe aqui detalhar.

Ao final da mensagem o símbolo de fim de mensagem deve ser transmitido seguido de tantos bits quantos forem necessários para garantir que a cadeia está no range de final de mensagem.

5.5 Dicionários

A compressão de dados baseada em dicionários parte deste princípio: grupos de símbolos consecutivos, as *frases*, estão vinculados a índices de algum dicionário e os índices ocupam menos espaço que as frases às quais estão vinculados. O dicionário é uma lista de frases que se espera sejam muito utilizadas. Possui uma implementação eficiente pois os índices podem ser criados no tamanho de palavras do computador, evitando-se assim a manipulação de bits e alinhamentos. Formalmente um dicionário $D = (M, C)$ é um conjunto finito de frases M e uma função C que associa cada mensagem de M a um determinado código.

Assim como nos algoritmos probabilísticos, a compressão por dicionário pode ser estática, semi-adaptativa ou adaptativa. *Fraseamento* é a tarefa de dividir o texto em frases. Uma das estratégias mais comuns para o fraseamento é o algoritmo guloso que encontra no texto a maior frase que pertence a M e a substitui por seu índice. Esta estratégia não é necessariamente ótima. A tarefa do fraseamento pode ser transformada no problema de encontrar o caminho de comprimento mínimo num grafo, que pode ser resolvido de forma eficiente. Embora não seja ótimo o fraseamento guloso é o mais utilizado.

Codificadores estáticos baseados em dicionários

O mais comum é o codificador de digramas e funciona da seguinte forma: Examinam-se os dois próximos símbolos a serem codificados. Se o digrama fizer parte do dicionário, este será substituído por sua representação. Caso contrário, grava-se o primeiro símbolo do digrama e se repete a operação até o final da mensagem. Este processo funciona bem na língua inglesa pois há

somente 96 caracteres de texto e os $(256 - 96)$ 160 possíveis combinações de um byte podem ser utilizados para representar o 160 digramas mais freqüentes do texto. Na língua portuguesa há menos eficiência. Usam-se as letras do alfabeto inglês K, W e Y em anglicismos e abreviações, há o $Ç$ e as letras acentuadas com trema, acento grave, acento agudo, acento circunflexo e til, que podem estar tanto em maiúscula como em minúscula. Eventualmente pode-se colocar no dicionário trigramas, tetragramas ou até pentagramas mais freqüentes como `‘‘ th’’` na língua inglesa, mas há maior perda de tempo no processo de compressão de dados. Este processo também pode ser ampliado para uso com dicionário estáticos de n-gramas, criando-se palavras com 4, 8 ou 12 bits de tamanho onde os primeiros 4 bits indicarão o tamanho da entrada do dicionário.

Codificadores semi-adaptativos baseados em dicionários

Em obras técnicas há expressões que ocorrem com bastante freqüência, como neste texto as palavras "entropia", "seqüência" e "DNA". Há também frases como "estimativa da entropia" candidatas a uma entrada no dicionário. A grande vantagem neste caso é que as frases normalmente são bem maiores que os índices utilizados para representá-las e com isto se pressupõe uma boa compressão de dados. O problema da determinação de um dicionário ótimo para um dado texto é NP-Difícil. O que tem sido feito na prática é preencher inicialmente o dicionário com os digramas mais freqüentes.

Quando algum trigrama for muito freqüente, ele é inserido. Neste ponto os digramas que o formam são analisados e se forem pouco freqüentes fora do trigrama estes são excluídos do dicionário. Por exemplo, sejam os digramas `‘‘ a’’` e `‘‘a ’’` muito freqüentes. Ao ser inserido o trigrama `‘‘ a ’’` analisam-se as ocorrências dos digramas anteriores fora deste, com a remoção dos digramas pouco freqüentes.

Codificadores adaptativos baseados em dicionários

Após o aparecimento do codificador Ziv-Lempel em 1977, a utilização de modelos adaptativos baseados em dicionários ficou basicamente reduzida ao codificador Ziv-Lempel e às suas diversas variações. Estes recebem o prefixo LZ em sua denominação. Os criados por Jacob Ziv e Abraham Lempel foram os LZ77 e LZ78, devido à sua criação nos anos de 1977 e 1978. Bell, Cleary e Witten [5] fazem referência a 12 métodos LZ. Os códigos de Ziv Lempel se caracterizam pela utilização do texto (mensagem) já codificado como dicionário para o restante da codificação.

O algoritmo é iterativo e recebe uma mensagem $S = \{s_1 s_2 \ldots s_n\}$ e devolve a codificação de S no alfabeto $\{0, 1\}$. Seja s a seqüência de símbolos de S ainda não codificados, $\overline{s}$ a cadeia de símbolos já codificada e s_q o que será codificado na próxima iteração; s_q é alguma subcadeia

que prefixa s. No início do processo, o dicionário está vazio, $\overline{s} = \{\}$, $s = S$ e o $s_q = \{s_1\}$. Partindo deste ponto, para cada s é examinada a subcadeia $\overline{s}.s_q$ para a determinação da maior subseqüência comum entre alguma subcadeia que prefixa s e alguma subcadeia que pertença a $\overline{s}.s_q$ cujo primeiro símbolo pertença a $\overline{s}$. Seja s_p esta subcadeia. Se s_p não for vazia, então $s_q = s_p$. Caso contrário, $s_q = \{s_1\}$, o primeiro símbolo de s. Após a codificação de s_q, são feitos os ajustes de s e $\overline{s}$ para a próxima iteração. $\overline{s} \leftarrow \overline{s}.s_q$, $s \leftarrow s - s_q$. O processo é repetido até que $s = \{\}$. A conversão da representação de $\overline{s}$ em números binários será omitida, por não ser relevante e é tratada exclusivamente como caracteres, para este texto.

Suponha uma pequena mensagem de poucos símbolos:

``caaaaaaaaaaaaaaaaaaaaaaaaatatatatatatatatatatatat''

[c(a)^{26}(ta)^{11}t]. Neste texto a compressão é iniciada com c. Na seqüência restante não há ocorrências anteriores, de forma que $s_q = \{\text{a}\}$, sendo agora $\overline{s} = \{\text{ca}\}$. A partir deste instante o processo detecta repetições, pois para os próximos 25 símbolos há uma repetição da seqüência definida a partir do segundo símbolo. Assim $\overline{s} = \{\text{ca}(2, 25)\}$ indicando que os próximos 25 símbolos (tamanho de s_p) repetem a seqüência iniciada na posição 2 do texto original, já codificado ($s_q = s_p$). Pode haver sobreposição entre a seqüência a codificar e o texto já codificado. O próximo símbolo a ser codificado é um t que ainda não ocorreu anteriormente (s_q é o primeiro símbolo de s). Desta forma codifica-se o t, ficando $\overline{s} = \{\text{ca}(2, 25)\text{t}\}$. Até aqui foram codificados 28 símbolos mas ainda faltam os 22 restantes. Mas estes repetem a seqüência iniciada na posição 27, ficando portanto a representação dos 50 símbolos originais do texto em $\overline{s} = \{\text{ca}(2, 25)\text{t}(27, 22)\}$. Este exemplo foi escolhido sob medida para o mecanismo da codificação dos algoritmos do tipo LZ. Embora a ocorrência de taxas de compressão desta eficiência sejam raras, o algoritmo é eficiente e funciona muito bem para a língua inglesa. A grande vantagem deste método é que não é necessário representar o dicionário pois este está no próprio texto codificado anteriormente seguido do texto a ser codificado. As variações dos códigos LZ estão no tamanho do texto anterior a ser examinado, no texto a frente a ser codificado e na estrutura para a busca da maior seqüência de símbolos que sejam comuns tanto ao dicionário quanto ao início do texto a ser codificado.

O termo *janela* designa a memória utilizada pelos codificadores LZ. Esta é dividida em dois segmentos contíguos: a *janela anterior* que contém os símbolos já codificados e seu restante para aqueles ainda não codificados.

LZ77

Este algoritmo examina a janela anterior (inicializada normalmente como espaços em branco) para determinar o tamanho e a posição da maior cadeia comum de símbolos entre esta e o que está

sendo codificado. São representados o deslocamento (quantos símbolos anteriores ao primeiro a codificar), o tamanho da cadeia e o símbolo seguinte à cadeia. É feito um deslocamento na janela do tamanho da codificação feita (tamanho da cadeia acrescido de um). Quando se examina a janela, pode haver sobreposição entre o que já foi codificado e o que falta codificar para se definir a maior cadeia que já ocorreu. Necessariamente deve haver um deslocamento positivo. Caso esta cadeia esteja vazia, costuma-se representar o próximo símbolo como (0, 0, ''x'') sendo ''x'' o símbolo seguinte a ser codificado. O primeiro zero indica o deslocamento e o segundo o tamanho da cadeia que coincidiu com o início da seqüência a codificar. As implementações variam no tamanho da janela, sendo usual o de 8192. São reservados de 10 a 20 símbolos para o que falta codificar e os demais são destinados à janela anterior. Para codificar a mensagem

''aataaataatataaa''

examinando 7 símbolos para trás e 5 à frente (tamanho de janela de 12 símbolos):

$$(0,0,\mathtt{a})(1,1,\mathtt{t})(3,2,\mathtt{a})(4,3,\mathtt{t})(5,4,\mathtt{a})$$

A seqüência final representada no computador utiliza 10 bytes, uma vez que tanto o deslocamento quanto o tamanho da cadeia podem ser representados em 4 bits. O LZ77 é muito utilizado; são baseados neste algoritmo PKZip, Zip, LHarc, PGN, gzip e ARJ.

LZ78

Este algoritmo cria um dicionário em tempo de execução. Cada entrada no dicionário consiste em alguma entrada anterior seguida de um símbolo. Seja d_i a i-ésima entrada no dicionário. O índice 0 denota a entrada vazia. Fica a codificação de:

''aataaataatataaa''

$$d_1 \leftarrow (0,\mathtt{a}), d_2 \leftarrow (1,\mathtt{t}), d_3 \leftarrow (1,\mathtt{a}), d_4 \leftarrow (2,\mathtt{a}), d_5 \leftarrow (4,\mathtt{t}), d_6 \leftarrow (3,\mathtt{a}).$$

A primeira entrada do dicionário é a, a segunda at, a terceira aa, a quarta ata, a quinta atat e a sexta aaa. Na prática não é possível criar um dicionário de tamanho indefinido. Chega-se em um certo ponto em que ele é reinicializado, como no começo do processo (vazio).

LZW

É uma variação importante do LZ78. Trata-se de inicializar o dicionário com os símbolos do alfabeto e a partir daí incrementar o dicionário com as novas frases, até que o mesmo esteja cheio, quando então passa a não mais receber frases. A vantagem deste método é que ele não precisa da transmissão de um símbolo a mais após cada apontador. Seja $\Sigma = \{a, t\}$ e a mensagem

$$S = \text{''aatatataaa''}.$$

O dicionário é inicializado com $d_0 \leftarrow$ a e $d_1 \leftarrow$ t. A saída deste exemplo fica:

001352.

Estes valores foram obtidos assim: os dois primeiros zeros correspondem aos símbolos pré inicializados a. Quando se lê o segundo a, cria-se uma nova entrada no dicionário $d_2 \leftarrow$ aa. O 1 é obtido do símbolo pré inicializado t e o dicionário recebe $d_3 \leftarrow$ at. O próximo símbolo 3 é obtido do último símbolo inserido no dicionário $d_3 \leftarrow$ at e o dicionário recebe $d_4 \leftarrow$ ta. O próximo símbolo d_5 possui uma característica peculiar: ele ainda não está definido no dicionário mas aparece na codificação. Sabe-se que da última entrada codificada na saída os dois primeiros símbolos de d_5 são at. Por convenção, quando uma entrada no dicionário não foi totalmente concluída, seu último símbolo é igual ao seu primeiro e neste caso $d_5 \leftarrow$ ata. O 2 final sai diretamente do dicionário e após esta codificação o dicionário recebe $d_6 \leftarrow$ ataa, mas a codificação já chegou ao final.

Como todos os outros métodos, este também tem suas vantagens e desvantagens, cabendo a quem vai efetivar a compressão escolher o método mais adequado às suas necessidades.

5.6 Custos da compressão

A estimativa da entropia está associada ao modelo e à distribuição de probabilidades obtidas da utilização deste modelo na mensagem. O compressor possui um custo a ele associado. Este custo é passado para a codificação da mensagem, utilizando mais bits que o necessário para representá-la. Em poucos casos específicos é possível que o compressor atinja exatamente a estimativa da entropia, mas no geral seu custo acaba aparecendo, em alguns casos, de forma significativa.

Um exemplo onde o custo do compressor é zero está na mensagem

``acagtctg'',

tanto para o código de Shannon-Fano quanto para o código de Huffman, supondo a utilização de um modelo de contextos finitos de ordem zero. Em ambos os casos cada símbolo seria representado com 2 bits por símbolo e como são equiprováveis a estimativa da entropia é exatamente esta.

Alguns custos são inevitáveis, por exemplo, o da gravação física do arquivo no disco. É necessário ao compressor a identificação do final da mensagem. A menor unidade que pode ser endereçada na memória é o byte. Caso a mensagem do exemplo acima possua um símbolo a mais, ela será representada por três bytes (24 bits para 9 símbolos) e de alguma forma o compressor necessitaria saber que a mensagem terminou. Apesar de que no último byte a maioria dos bits

não fazem parte da mensagem, ele está fisicamente vinculado ao arquivo comprimido e os 6 bits finais do arquivo não podem sofrer a decodificação. Há diversas formas no trato deste custo, mas ele normalmente é pequeno, pois uma mensagem possui milhares de símbolos e o último byte somente 8 bits.

Outro custo inevitável, para os modelos semi-adaptativos, é o envio do livro código. Se a ordem do modelo for bem escolhida, o custo do livro código é pequeno e a mensagem comprimida (livro código mais mensagem) é normalmente menor que a utilização de modelos estáticos ou adaptativos. Porém a representação do livro código não faz parte da estimativa da entropia, sendo desta forma um custo.

Também é inevitável o custo da manutenção dos modelos adaptativos. Um exemplo está na mensagem

``cagtacgt''.

No início há uma representação equiprovável para os quatro símbolos. Ao aparecer o primeiro símbolo c, este ganha um peso maior e sua representação se altera. Na representação do próximo símbolo a, sua probabilidade está menor o que provoca um custo adicional em sua representação, o mesmo ocorrendo até o quarto símbolo da mensagem. Para a representação do quinto símbolo as probabilidades ficaram novamente 1/4 para cada símbolo, mas do sexto até o final da mensagem há um custo adicional na representação dos restantes. Embora os modelos adaptativos sejam muito utilizados, seu custo acaba sendo passado para a representação da mensagem comprimida. Desta forma, a mensagem comprimida utiliza mais bits que sua estimativa de entropia prevê.

Além dos custos inevitáveis há os custos vinculados ao processo de compressão. Não serão aqui descritos todos os detalhes do custo de cada tipo de compressor. Dois serão destacados: o do compressor aritmético e o do método LZ.

Ao ser utilizada a compressão aritmética, há sempre a probabilidade do final da mensagem. Embora esta somente seja codificada ao fim da mensagem, para cada símbolo codificado há uma probabilidade vinculada ao fim da mensagem. Este fato em média diminui a probabilidade dos demais símbolos do alfabeto. Desta forma a possibilidade da representação do final da mensagem é um custo que acompanha toda a codificação da mesma. Embora haja outros custos envolvidos no processo da compressão aritmética, a citação de apenas este já demonstra que, embora muito eficiente, o compressor aritmético também supera a estimativa da entropia para uma mensagem, utilizando algum modelo probabilístico.

Nos compressores que utilizam algum dos métodos LZ, ao ser representada uma mensagem o tamanho da memória utilizada é definido. Ao ser definida uma memória de 4096 bytes são necessários 12 bits para representar cada entrada no dicionário. Quando se define o tamanho da seqüência a codificar, também se define um número de bits necessários para representá-la.

Para uma seqüência de tamanho 16 são necessários 4 bits em sua representação. Desta forma são necessários $12 + 4 = 16$ bits para codificar cada seqüência selecionada da mensagem. Por exemplo, a codificação de símbolos de nucleotídeos que constituem uma cadeia de DNA. Caso a cadeia polinucleotídica possua um tamanho pequeno que coincida com a memória passada, por exemplo 7 nucleotídeos, sua codificação necessitaria de 16 bits. É mais do que o previsto pela distribuição aleatória para as cadeias de DNA (dois bits por símbolo).

Quando se faz a estimativa da entropia de uma mensagem, se determina um valor ótimo, que dificilmente é atingível por um compressor, embora teoricamente seja possível, que os compressores estáticos atinjam este valor. O número de bits por símbolo obtidos em um compressor normalmente é maior que a entropia. Estimadores de entropia podem obter resultados menores que os compressores em geral para as estimativas da entropia. Desta forma, a utilização de estimadores de entropia normalmente refletem mais a realidade da entropia da mensagem que a utilização de compressores.

CAPÍTULO 6

Entropia e o Limite Íntron/Exon

Farach, Noordewier, Savari, Shepp, Wyner e Ziv [19] destacam as aplicações práticas do estudo da estimativa da entropia da seqüência primária do DNA para a localização de regiões com significado biológico. Tratam precisamente do problema da determinação dos limites entre os íntrons e os exons. Como a região dos exons normalmente é muito menor que a dos íntrons nos seres eucarióticos, os métodos utilizados para a estimativa da entropia destas regiões devem convergir rapidamente. Eles foram os primeiros a mostrar, por experiências, que as medidas de entropias são distintas para estas regiões. Foi utilizada a entropia para o algoritmo LZ (dicionário). Desta forma a entropia sai fora do contexto probabilístico onde está ligada ao modelo e entra no âmbito dos dicionários, vinculada ao fraseamento (seção 5.5).

6.1 O limite íntron/exon

DNA e proteínas são polímeros biológicos constituídos por nucleotídeos e aminoácidos, respectivamente. O gene é uma certa porção contígua do DNA que serve de base para a codificação de uma proteína. Nos seres eucarióticos o gene está dividido em intervalos intercalados de íntrons e exons. Porém somente a região dos exons é utilizada na codificação da proteína (capítulo 2). O problema da determinação dos limites íntrons/exons consiste na identificação do ponto exato em que começa e termina cada íntron e cada exon. Os íntrons quase sempre começam com gt e terminam com ag. Em resumo: na determinação dos gt/ag que significam início/fim de íntron dentre os gt/ag quaisquer da seqüência do DNA.

Foram estudados somente genes humanos com codificação completa, sendo 659 íntrons e 669 exons. O tamanho médio dos exons foi 184 bases e dos íntrons 867, com suas medianas com 139 e 434 bases e seus desvios padrões 96 e 583, respectivamente.

6.2 Métodos para a estimativa da entropia

Foi criado o algoritmo Match Length Entropy Estimator, que possui uma taxa de convergência relativamente rápida, para medir a estimativa da entropia de cadeias de DNA. As bases matemáticas destacam a entropia, o algoritmo LZ (seção 5.5) e as hipóteses dos estimadores de entropia. Como o comprimento dos exons é bem pequeno, o tamanho da janela do algoritmo LZ deve ser diminuta, para evitar perda de dados. A entropia foi escolhida porque é uma medida natural de complexidade, compressibilidade, previsibilidade e aleatoriedade.

Seja Σ o alfabeto e λ seu tamanho. Seja $(X_1, X_2, \ldots)$ uma seqüência de variáveis aleatórias com probabilidades $\mathcal{P}$. Ou seja, para cada inteiro positivo l e cada possível seqüência x_1^l pertence a Σ^l com probabilidade $P(x_1^l) = P\{X_1^l = x_1^l\}$. Logo P associa cada seqüência à sua probabilidade em $\mathcal{P}$. A entropia é definida como o seguinte limite:

$$H(P) = \lim_{n\to\infty} \frac{E - [\log_2 P(X_1^n)]}{n} \tag{6.1}$$

onde $E[Y]$ indica o valor esperado de Y. A entropia é importante porque é uma propriedade da distribuição de probabilidades e não uma propriedade direta da seqüência de símbolos (seção 3.4). É possível expandir a noção da entropia para a memória.

Sejam U e V variáveis aleatórias e $P(u|v)$ a probabilidade de $U = u$ dado que $V = v$. A entropia condicional (seção 3.4) de U com relação a $V = v$ é definida por

$$H(U|v) = -\sum_{u\in U} P(u|v) \log_2 P(u|v)$$

e a entropia condicional de U em relação a V é definida como

$$H(U|V) = E_v[H(U|v)]$$

sendo E_v o valor esperado dos elementos de V.

Técnicas de estimativa da entropia

A técnica mais objetiva para a estimativa da entropia é o cálculo direto do valor esperado da distribuição de probabilidades, através da equação (6.1). Normalmente não há dados suficientes para ser obtida uma estimativa correta da entropia, mas se a seqüência possui uma baixa ordem do processo de Markov (seção 3.2) os resultados serão satisfatórios.

Outra prática comum para a estimativa da entropia é a utilização de compressores que lhe fornecem um limite superior (seção 3.4). Embora melhor[1] que a opção anterior, esta prática possui uma taxa de convergência muito baixa.

[1] Quando o compressor é universal, a taxa de compressão se aproxima da entropia se a mensagem for suficientemente longa.

O algoritmo LZ pode servir como estimador de entropia. Sua técnica de fraseamento consiste em iniciar o dicionário vazio. Cada nova entrada no dicionário é feita a partir de alguma entrada já existente seguida por um símbolo. Um exemplo está nos dados binários ($\lambda = 2$ e $\Sigma = \{0,1\}$) 0010111000101 cujo fraseamento pode ser feito através das vírgulas adicionadas $\{0, 01, 011, 1, 00, 010, \ldots\}$ Seja C_n o número de entradas no dicionário. Ziv e Lempel [59] provaram que a quantidade $\frac{C_n \log C_n}{n} \to H$ se $n \to \infty$.

$$\frac{C_n \log C_n}{n} \to H \tag{6.2}$$

É um candidato natural para a estimativa da entropia. Embora universal e prático, este método para estimar a entropia possui uma taxa de convergência notoriamente baixa, principalmente quando a fonte não é estacionária (seção 3.2).

Uma versão modificada dos algoritmos LZ também pode servir para estimar a entropia. Uma de suas abordagens alternativas está na utilização de uma janela fixa e da maior subcadeia. Seja X_1^∞ a seqüência e w o tamanho da janela. Pode-se se representar a janela como X_{-w+1}^0. Seja L o tamanho da maior subseqüência, a partir do início da seqüência, que exista na janela. Matematicamente, a seqüência pode ser descrita através da expressão $X_1^L \subseteq X_{-w+1}^0$ e $X_1^{L+1} \nsubseteq X_{-w+1}^0$, na qual $\subseteq$ é uma subcadeia da janela e $\nsubseteq$ necessariamente não é uma subcadeia. O estimador de entropia varre a fonte símbolo a símbolo. Desta forma a janela é X_{i-w+1}^i e a maior subseqüência é $X_{i+1}^{i+L_i}$. Seja $\overline{L}$ a média das maiores subcadeias L_i. O estimador de entropia neste caso é definido pela expressão.

$$\widehat{H} = \frac{\log_2 w}{\overline{L}} \tag{6.3}$$

Este estimador possui um erro de $O\left(\frac{1}{\log w}\right)$. As principais fontes do erro são a escolha de w e o desconhecimento do tamanho da memória no processo. Este estimador foi chamado de Sliding Window Entropy Estimate.

6.3 Análise do estimador de entropia

O Slinding Window Entropy Estimate foi criado especialmente para tentar medir a estimativa de entropia do DNA; sua principal característica é a rápida convergência. Enquanto os algoritmos LZ tradicionais fazem uma medição por entrada do dicionário, este faz uma por símbolo da mensagem, o que garante sua taxa de convergência.

Foram supostas algumas condições na variação do estimador de entropia. Há razões para acreditar que, em muitos casos, o erro teórico será bem pequeno. A estimativa da entropia é apenas aproximada, pois a memória do processo pode ser maior que w. O DNA não é um processo estacionário (seção 3.2) e processos não estacionários não podem ser caracterizados

pela entropia. O DNA não é um processo aleatório. A matemática serve como uma forma de abordagem das seqüências de DNA; não há pretensão de definir a entropia para o DNA.

6.4 Resultados e conclusões

A determinação do início dos íntrons foi relativamente bem sucedida mas o mesmo não ocorreu com a determinação do término. Desta forma, não foram identificados padrões indicativos para o final dos íntrons, o qual é mais difícil de detectar do que o início.

Na determinação do início dos íntrons foram utilizadas 39439 cadeias iniciadas por `gt`; 579 iniciavam íntrons. Seja R o conjunto das seqüências iniciadas pelos pares `gt` da base de dados adotada e seja k o número de bases à frente do par `gt` que utilizado para a estimativa da entropia. $H(R_k|R_1^{k-1}) \leq H(R_n) \leq 2$, sendo considerado $H(R_k|R_1^{k-1}) \approx 2$ se $H(R_k|R_1^{k-1}) > 1,8$. Como há 4 bases na representação da estrutura primária do DNA, foi arbitrado o exame de até 5 bases à frente do par `gt`, pois os dados seriam insuficientes para uma base estatística confiável se $k > 5$ ($4^k \geq 4^6 = 4096$). Aplicando o estimador de entropia em todos os pares `gt` $H(R_k|R_1^{k-1}) \approx 2$ para todo $k < 31$ foi obtido. Houve um resultado similar para experiências simétricas à esquerda dos pares `gt`.

Os resultados quando aplicados somente aos pares `gt` que iniciavam os íntrons foram similares para $k > 4$ (sendo para $k = 4$ $H(R) = 1,74$) à direita dos pares `gt` e para $k > 3$ à esquerda, gerando um padrão da forma `xxxgtxxxx`. A partir daí, foi reduzido o espaço para 7−tuplas ($4^7 = 16384$ possibilidades) em volta dos pares `gt`, sobre as quais foi utilizado o teste estatístico de Neyman-Pearson. Feito o teste em todos os pares `gt` foram encontrados os padrões

`aaggtaagt`

`aaggtgagt`

`gacgtaagt`

`gacgtgagt`

para o início dos íntrons. A figura 6.1 mostra a entropia condicional aplicada a todos os pares GT (na forma `xxxxxgtxxxxx`)e somente aos pares GT que constituem início de íntron (na forma `xxxgtxxxx`).

A principal conclusão destacou a estimativa de entropia entre os íntrons e os exons. A entropia dos íntrons geralmente é menor que a dos exons, o que é de certa forma surpreendente. Os íntrons aparentemente estão mais sujeitos aos mecanismos de mutações aleatórias sucessivas, sem necessariamente definir um gene novo. A explicação encontrada foi que uma grande parte dos

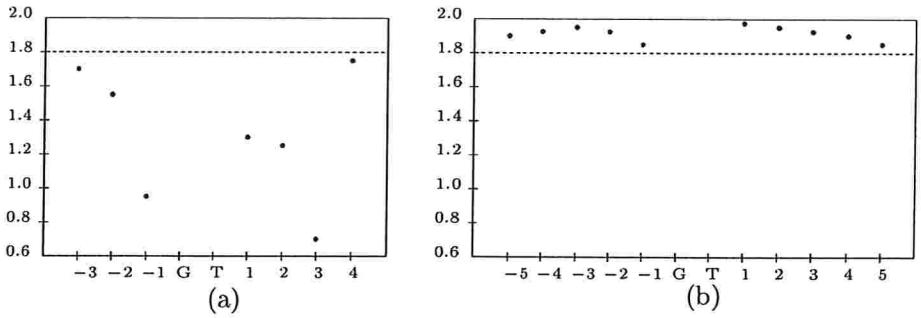


Figura 6.1: Entropia condicional em torno dos pares GT

íntrons está sujeito a forças restauradoras. Algumas destas partes podem servir para determinar os limites entre os íntrons e exons.

Embora a entropia seja uma medida da informação, a maioria dos estimadores de entropia falharam quando aplicados à estrutura primária do DNA, provavelmente por não convergirem rapidamente. Acredita-se que os pesquisadores de algoritmo deverão encontrar um estimador de entropia inovador que resolva de forma assintótica o problema da determinação dos limites entre os íntrons e os exons.

Capítulo 7

Entropia Utilizando Comparações Inexatas

Loewenstern e Yianilos [31] questionam o conceito de estimativa da entropia do DNA e criam um estimador de entropia, o CDNA. A partir deste, um compressor (o compressor CDNA) que reduziu bastante o valor desta estimativa em relação aos compressores existentes. Expandiram os modelos de contexto finito através de comparações inexatas e aprimoraram a estimativa da entropia nas seqüências do DNA de diversos tipos de seres, inclusive genes humanos.

7.1 Estimativas de entropia

Se a estrutura primária do DNA fosse aleatória, bastariam dois bits por nucleotídeo para representá-la, pois o alfabeto $\Sigma = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ possui tamanho quatro. Espera-se do DNA uma pequena aleatoriedade, mas surpreendentemente os compressores tradicionais têm obtido baixas taxas de compressão nos diversos genomas.

Os modelos de contexto finito convencionais focam os nucleotídeos imediatamente anteriores ao que se deseja representar. Contextos maiores aprimoram o modelo, mas aumentam a possibilidade do contexto não ter ocorrido anteriormente (seção 4.1). Isto levou ao desenvolvimento de modelos de contextos de diversas ordens, o blending (seção 4.6). Estes utilizam contextos maiores que são reduzidos quando ocorre o problema da freqüência zero. Porém nas cadeias de DNA os contextos variáveis não surtiram o efeito esperado, exceto nos raros casos de longas repetições nos polinucleotídeos. Foi proposta a comparação inexata como uma tentativa de melhorar a estimativa da entropia do DNA. Esta tornou possível a utilização de contextos maiores e a adoção da distância de Hamming como medida das divergências de comparação.

Conceitualmente a entropia somente poderia ser utilizada nas seqüências genômicas se o

modelo fosse compatível com o processo do DNA[1], se este processo fosse bem comportado[2] e se a cadeia fosse infinita. Como tais condições não ocorrem[3], são utilizados compressores para obter a estimativa da entropia do DNA. A estimativa é obtida dividindo o tamanho da seqüência comprimida pelo número de bases da mensagem. Este processo superestima a estimativa da entropia. Para as cadeias do DNA esta fica em torno de 1,9 bits/nucleotídeo. O compressor do UNIX[4], que é baseado nos algoritmos LZ (seção 5.5), foi utilizado no Humretblas[5] e obteve 2,14 bits por caractere, ficando pior que a estimativa aleatória. Existe a possibilidade de que a estimativa seja atenuada com a utilização somente da parte final do arquivo. No caso da parte final da seqüência possuir uma entropia menor, ocorreria a descaracterização da estimativa. A solução utilizada foi chamada de validação cruzada[6] e está na divisão da mensagens em um certo número de segmentos de um mesmo tamanho. A estimativa da entropia é feita da seguinte forma: cada segmento é considerado como se fosse o último e são utilizados os demais segmentos para calibrar o estimador. A média das estimativas individuais é tomada como a estimativa da entropia da mensagem.

7.2 As bases do CDNA

Seja S_1 e S_2 cadeias de mesmo tamanho que serão comparadas entre si. A *distância de Hamming* $h(S_1, S_2)$ indica o número de posições divergentes entre S_1 e S_2. Se $h(S_1, S_2) = 0$ então $S_1 = S_2$ e se $h(S_1, S_2) = 3$ há 3 divergências entre S_1 e S_2, não importando a posição destas.

A tabela 1 mostra a aplicação da fórmula $|T| \cdot 3^h/4^{|S|} \cdot \binom{|S|}{h}$. Esta representa a quantidade de ocorrências esperadas para uma dada distância de Hamming no segmento $|T|$. $|T|$ representa a cadeia na qual se comparam as subcadeias, $|S|$ é o tamanho das cadeias $S_1 \in S$ e $S_2 \in T$ e h a distância de Hamming. Para fazer a tabela 1 foi utilizado $0 \leq h \leq 20$, $|S| = 20$ e $|T| = 7/8 \cdot 180.388$ e o 7/8 saiu da divisão do gene Humretblas em 8 partes iguais, das quais cada uma foi utilizada para gerar as cadeias S_1 e as demais concatenadas para gerar as cadeias S_2. S_1 e S_2 neste caso são cada segmento contíguo do gene Humretblas de tamanho 20, respectivamente no 1/8 do gene ou nos 7/8 restantes.

Nesta tabela observa-se que para pequenos valores de h, entre 0 e 5, o número de ocorrências supera o esperado e para maiores valores, entre 15 e 20, a quantidade de ocorrências é menor

[1] Modelo gerador (seção 3.2) do DNA.

[2] Ergódico e estacionário (seções 3.2 e 4.3).

[3] Não se conhece o processo do DNA, o DNA não é estacionário (embora seja ergódico) e as cadeias possuem um tamanho reduzido.

[4] Marca registrada da X/Open Company Ltd.

[5] Gene humano da susceptibilidade ao retinoblastoma com 180.388 bases.

[6] *Cross validation* no original.

que a esperada, o que indica um baixo índice de aleatoriedade das cadeias de DNA. Também nota-se que para baixos valores de h a possibilidade de acerto do próximo caractere é bem maior que a utilizada na distribuição aleatória.

TABELA 1: ACERTOS VERSUS PREVISÕES (GENE HUMRETBLAS JANELA DE 20)

Distância de Hamming	Número esperado de ocorrências	Número observado de ocorrências	Percentagem com pelo menos uma ocorrência	Percentagem de acerto do próximo símbolo
0	$1{,}44 \times 10^{-7}$	0,387	4,4	90,91
1	$8{,}61 \times 10^{-6}$	0,143	7,2	83,77
2	$2{,}45 \times 10^{-4}$	0,278	11,0	78,19
3	$4{,}42 \times 10^{-3}$	0,505	17,6	66,00
4	0,0563	1,184	33,2	45,91
5	0,541	4,148	78,2	35,86
6	4,01	18,759	98,3	31,90
7	24,3	80,69	100	29,19
8	119	304,9	100	28,35
9	475	986,0	100	28,09
10	1.566	2.713	100	27,86
11	4.271	6.286	100	27,68
12	9.609	12.300	100	27,53
13	17.740	20.172	100	27,33
14	26.610	27.387	100	27,09
15	31.932	30.356	100	26,86
16	29.936	26.813	100	26,57
17	21.131	18.200	100	26,28
18	10.566	8.931	100	25,94
19	3.337	2.826	100	25,69
20	500	434,5	100	25,43

Seja Σ um alfabeto, σ, b símbolos deste alfabeto, $S, T \in \Sigma^*$, $S[i]$ o i-ésimo símbolo de S, $S[i,j]$ do i-ésimo ao j-ésimo símbolos de S, $S \cdot T$ a concatenação de S e T, $S : b$ a concatenação de S com o símbolo b, $\mathtt{suf}(S,i)$ os primeiros i caracteres de S sendo $0 \leq i \leq |S|$, $\mathtt{conf}(S,T,h)$ o conjunto de todas as subcadeias T tal que S e T tenham exatamente distância de Hamming h. A probabilidade $\Pr_{w,h}$ representa a probabilidade do próximo símbolo de S.

$$\Pr_{w,h}(b|S,T) = \frac{1+|\mathtt{conf}(\mathtt{suf}(S,w):b,T,h)|}{\sum_{\sigma\in\Sigma} 1+|\mathtt{conf}(\mathtt{suf}(S,w):\sigma,T,h)|}$$

w assume algum valor de algum conjunto W. Uma vez definido w é esperado que quanto maior h maior será o número de ocorrências de S em T.

Observe que se cada termo $\mathtt{conf}(\mathtt{suf}(S,w) : \sigma, T, h)$ for o conjunto vazio, então $Pr_{w,h}$ atribui para cada b em Σ a probabilidade uniforme $1/|\Sigma|$. Serão ignoradas todas as distâncias de

Hamming para as quais $\texttt{conf}(\texttt{suf}(S,w),T[1,|T|-1],h)$ for vazio, sendo relevantes somente as demais.

Fixadas as cadeias S e T e um comprimento w denotamos por f o menor valor de h tal que em $T[1,|T|-1]$ ocorra uma cadeia de tamanho w que possua uma distância de Hamming de $\texttt{suf}(S,w)$ exatamente h, ou seja: f a primeira distância de Hamming relevante. Seja $F(f,w,S,T)$ uma função que somente assuma os valores um ou zero. Assume um se f for o menor valor possível de h tal que $\texttt{conf}(\texttt{suf}(S,w),T[1,|T|-1],h) \neq \emptyset$ e zero caso contrário. Neste caso $0 \leq f \leq w$ e $h \geq f$.

Uma possibilidade para determinar os valores de w e h está na colocação de pesos para estes valores. Sejam Ψ e Υ respectivamente os pesos para os valores possíveis de w e h ($\sum_{w \in W} \psi_w = 1$ e $\sum_{h=f}^{w} \upsilon_{w.f.h} = 1$). Cabe destacar que, uma vez escolhido w, o valor de f está determinado e $f \leq h \leq w$. Os pesos para h devem ser zero se $h < f$ ou $h > w$. O algoritmo CDNA se baseia na otimização dos pesos Ψ e Υ através do algoritmo Expectation Maximization (Baum-Welch/EM) para achar o $\texttt{arg}\ \texttt{max}_{\Psi,\Upsilon} P(S|T,\Psi,\Upsilon)$, a partir de uma distribuição uniforme. Também foi criado o compressor CDNA com o qual foram obtidos dados práticos que foram comparados com os de outros compressores. Destaca-se que a estimativa de entropia do CDNA foi sempre melhor (menor) que o do compressor CDNA.

7.3 Os dados experimentais e os resultados obtidos

Foram utilizados genes de diversas espécies. Em especial 13 de mamíferos (dentre os quais o Humretblas da tabela 1), um íntron extenso de mamífero (Humdystrop), de procariotos, plantas, fermentos, etc.

Os algoritmos CDNA e compressor CDNA foram utilizados para fazer a estimativa da entropia do DNA e seus resultados foram comparados com o compressor do UNIX, com os algoritmos H_1, H_4, H_6[7] e o Biocompress-2, algoritmo de Grumbach e Tahi de 1994 [25]. O compressor do UNIX invariavelmente ficou pior que a distribuição aleatória (dois bits por símbolo) para a estimativa da entropia do DNA, provavelmente por limitações no tamanho do dicionário (impedindo que houvesse convergência). Os compressores probabilísticos convencionais ficaram com a estimativa de entropia em torno de 1,95, ficando menores que 1,9 somente para poucos casos. A utilização do H_6 raramente melhorou a estimativa da entropia do DNA com relação ao H_4, provavelmente por restrições no tamanho das cadeias. Os resultados do Biocompress-2 normalmente foram mais efetivos (menores) que os compressores probabilísticos, para cadeias de DNA, mas tanto o CDNA quanto o compressor CDNA em geral superaram o Biocompress-2, obtendo os menores valores de estimativa da entropia do DNA.

[7] Estimadores probabilísticos convencionais de entropia, para 1, 4 e 6 bases antecessoras respectivamente.

O gene Humretblas foi utilizado para fazer diversas medidas no CDNA. A primeira delas que cabe destaque é o número de iterações EM (expectation maximization). Foram utilizadas 200 iterações, mas geralmente há convergência após 150. Para reduzir a exigência de memória foram utilizadas janelas de tamanhos pares de 2 a 24. Janelas maiores que 25 normalmente sobrecarregam o CDNA, pois o número de parâmetros tem crescimento proporcional a w^2 onde w é o tamanho da janela. A utilização de janelas de tamanhos diferentes normalmente melhora a estimativa da entropia, mas o benefício constatado foi pequeno. Foram utilizadas cadeias com 5.000 nucleotídeos para treino, pois tamanhos maiores normalmente resultavam ganhos insignificantes e há uma notável convergência entre os valores de treino e teste.

Dois experimentos em especial merecem destaque no que se refere à estimativa da entropia do DNA. Eles reportam a estimativas de entropia cruzadas. O primeiro treina o CDNA em exons e testa em íntrons e vice versa, tanto para mamíferos quanto para fermentos. A estimativa de entropia cruzada ficou acima de $1,9$. O segundo treina em íntrons de mamíferos e testa nos íntrons de fermento e vice versa, fazendo o mesmo com relação à região dos exons. Novamente os resultados ficaram acima de $1,9$.

Os exons em geral possuem menor taxa de compressão (seção 3.3) que os íntrons. Foi utilizado o CDNA em exons de regiões não redundantes de 490 genes humanos. Foi observado que a estimativa de entropia dos exons era maior que a do gene inteiro, conforme previsto. Não houve ganho significativo na estimativa da entropia do DNA dos exons, mas o CDNA apresentou uma pequena melhora (diminuição) de seu valor.

7.4 Considerações finais e conclusões

Há uma questão sobre qual estimativa de entropia deve ser esperada para cadeias do DNA e por que. Os modelos probabilísticos não revelam muito da natureza do DNA[8]. Provavelmente entidades similares a palavras e frases sejam inexpressivas nos genomas, mesmo que a exigência de comparações exatas sejam relaxadas.

A degeneração do código genético pode contribuir para a pouca compressibilidade das cadeias do DNA. Não se sabe quantas seqüências diferentes são equivalentes do ponto de vista da evolução natural. Pequenas mutações alteram pouco a mensagem do DNA como um todo.

O CDNA utiliza somente um domínio do conhecimento do DNA: as seqüências do DNA possuem mais repetições próximas que a distribuição aleatória. É possível que utilizando outros domínios do conhecimento do DNA seja possível melhorar a estimativa da entropia dos exons.

[8]Embora o artigo não explicite neste ponto, podemos notar que o compressor do UNIX obteve invariavelmente estimativas de entropia maiores que 2 para cadeias de DNA. De alguma forma os algoritmos LZ também não conseguem capturar a essência do DNA.

Porém é provável que esta estimativa não seja muito menor que a obtida pelo CDNA. O tamanho dos exons também interfere em sua previsibilidade. Menos de um terço dos genes não redundantes utilizados possuíam exons com menos de 1000 bases.

A inclusão de comparações inexatas implicou na diminuição substancial na estimativa da entropia do DNA obtida pelo CDNA. Foram utilizadas diversas experiências como várias distâncias de Hamming para o trabalho com o complemento de seqüências invertidas. O próximo passo pode estar na utilização não apenas da distância de Hamming mas também do conhecimento de onde as diferenças ocorrem na comparação das seqüências.

Capítulo 8

Compressão de Seqüências Protéicas

Nevill-Manning e Witten [39] fazem um estudo na estrutura primária das proteínas, levando em conta as probabilidades de mutação dos aminoácidos num período de 62 milhões de anos. É feita uma generalização do blending a partir da qual é criado o compressor genérico CP (Compressor de Proteínas).

8.1 Proteína e DNA: as bases do compressor CP

Algumas características são destacadas tanto no DNA quanto nas proteínas (capítulo 2). O alfabeto do DNA possui 4 símbolos, enquanto o das proteínas 20. As bases nitrogenadas em si não possuem grande significado e estão sujeitas a mutações mais uniformes que os aminoácidos. Estes variam consideravelmente de tamanho, forma, estrutura e afinidade com a água. Podem ser neutros ou possuírem alguma carga elétrica. Esta variedade permite que as proteínas tenham uma grande diversidade de formas e dimensões, além de um amplo escopo de funções. A mutação de um aminoácido depende não só de suas características como também da funcionalidade obtida pela proteína alterada. Em alguns casos a mutação genética de um aminoácido pode levar à morte do organismo ou mesmo impedir que esta mutação seja passada aos seus descendentes[1].

As probabilidades de mutação genética de um aminoácido em outro[2] dependem de diversos fatores e a tabela 3 mostra estas probabilidades (em porcentagens) de mutação de um aminoácido em outro no período de 62 milhões de anos[3]. Cada linha da tabela soma 100 (por cento). A probabilidade é do aminoácido que consta na linha sofrer mutação para seu equivalente na coluna

[1] São citadas como causas possíveis da mutação genética no indivíduo: erro na cópia do DNA ou raio cósmico danificador.

[2] Há possibilidade não só de uma mutação transformar um aminoácido em um códon de fim de codificação como também haver inserção ou exclusão de nucleotídeos que modificariam completamente a proteína que está sendo sintetizada [58].

[3] Não há nenhuma referência à origem desta tabela no artigo.

$P(A \rightarrow S) = 10\%$ e $P(S \rightarrow A) = 12\%$.

Tabela 3 Probabilidade da substituição de aminoácidos nas proteínas válidas.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	17	4	3	3	3	3	5	9	2	5	7	5	2	3	3	10	6	1	2	8
R	5	21	5	4	1	6	6	4	3	3	6	15	2	2	2	5	4	1	2	4
N	5	5	19	10	1	4	6	8	4	3	4	6	1	2	2	8	6	0	2	3
D	5	4	9	25	1	4	11	6	2	3	4	6	1	2	3	7	4	0	1	3
C	8	2	2	2	32	2	2	4	1	6	8	3	2	3	2	6	5	1	2	7
Q	6	8	5	5	1	12	12	4	3	3	5	10	2	2	3	6	5	1	2	4
E	6	6	5	11	1	8	17	4	3	3	4	9	1	2	3	6	4	1	2	4
G	11	3	5	5	1	2	4	34	2	3	4	5	1	2	2	7	4	1	2	3
H	5	6	7	4	1	5	6	4	22	3	5	6	2	4	2	5	3	1	7	3
I	5	2	2	2	2	2	2	2	1	16	19	3	4	5	2	3	5	1	2	20
L	6	3	2	2	2	2	2	3	1	14	23	3	6	7	2	3	4	1	3	12
K	7	12	5	5	1	6	8	5	2	3	5	16	2	2	3	6	5	1	2	4
M	6	3	2	2	2	3	3	3	2	11	21	4	9	5	2	4	4	1	2	10
F	4	2	2	2	1	1	2	3	2	8	14	2	3	24	1	3	3	2	11	7
P	7	3	3	4	1	3	5	5	2	3	5	5	1	2	33	6	5	0	2	4
S	12	4	6	6	2	4	6	8	2	3	5	6	2	2	3	12	9	1	2	5
T	8	4	5	4	2	3	5	5	2	6	7	5	2	3	3	11	14	1	2	8
W	4	3	2	2	1	2	3	4	2	4	7	3	2	9	1	3	3	33	9	4
V	5	3	3	2	1	2	3	3	6	5	8	4	2	16	2	4	3	3	19	6
Y	8	3	2	2	2	2	3	3	1	19	15	3	4	4	2	4	6	1	2	16

Diversas espécies podem ter passados por processos evolutivos similares, possuindo assim proteínas com funções equivalentes. Normalmente estas proteínas possuem grande similaridade, o que acarreta um alto grau de redundância. Para garantir a confiabilidade do processo de compressão, o algoritmo CP foi utilizado em diversas proteínas de um mesmo organismo. Desta forma o processo busca identificar estruturas com significado biológico nas proteínas. Outra característica do DNA está na proporção de nucleotídeos utilizados na síntese protéica. Há uma grande variação de espécie para espécie, enquanto no E. Coli quase todo o DNA está envolvido na produção das proteínas na Fritillaria apenas 0,02% de seu código participa deste processo.

8.2 Compressor CP

Seja Σ um alfabeto, λ seu tamanho, $\sigma \in \Sigma$ um símbolo deste alfabeto, $P(\sigma)$ a probabilidade de ocorrer o símbolo σ e m a ordem máxima a ser utilizada. A técnica de blending (seção 4.6) utiliza a transmissão de cada símbolo σ utilizando o tamanho $\log_2 P(\sigma)$ bits, processando da esquerda para a direita na seqüência da mensagem. Por exemplo, se $m = 3$ e neste ponto da

mensagem os últimos três caracteres lidos foram `cgt` caso o contexto `cgt` tenha ocorrido dez vezes nas quais os símbolos `a`, `c`, `g`, `t` tenham sucedido o contexto respectivamente 2, 1, 5 e 2 vezes e se o sucessor for `t`, este será representado por $-\log_2 0,2 \approx 2,3$ bits. Caso `t` nunca houvesse aparecido, a técnica de blending iria reduzir o contexto para `gt` e assim sucessivas vezes até chegar em uma distribuição uniforme para os símbolos de Σ.

O principal motivo da utilização da técnica de blending é que os dados normalmente são escassos e mal distribuídos. Se a mensagem fosse ergódica (seção 3.2) e suficientemente grande, a utilização de um contexto de tamanho fixo seria aceitável. Citando as proteínas como exemplo, onde $\lambda = 20$ e se for utilizado $m = 5$ e 1000 amostras por contexto, seria necessária uma seqüência com tamanho em torno de 3.000.000.000 para treinar um algoritmo de contexto finito. Como normalmente as proteínas são bem menores que isto e mal distribuídas[4], a utilização de blending aprimora o modelo.

Ao invés de variar o tamanho do contexto, o algoritmo CP utiliza todos os contextos até um tamanho máximo m, com pesos variando de acordo com a similaridade do contexto em relação à seqüência na qual se deseja codificar o próximo símbolo. Para seqüências de tamanho 1 o peso do contexto é a probabilidade de seu símbolo sofrer mutação para o símbolo da seqüência corrente (conforme a tabela 3). Para contextos maiores as probabilidades de mutação são multiplicadas de forma que, se a seqüência sendo comprimida é `II` e o contexto é `HI`, o peso do contexto é $0,01 \cdot 0,16 = 0,0016$ ou $0,16\%$. Ao se utilizar todos os contextos de um mesmo tamanho, a previsibilidade ganha o suporte de mais dados. Símbolos similares ocorrem em contextos similares e as probabilidades para determinado tamanho englobam a dos tamanhos menores, pois todos os contextos menores são sufixados (limitado ao que aparece na mensagem) para atingir o tamanho dos contextos maiores. Desta forma o algoritmo CP utiliza todos os contextos de tamanho m na previsibilidade do próximo símbolo.

8.3 Resultados e conclusões

Foram utilizados os códigos genéticos de 4 seres. Uma bactéria que provoca doenças respiratórias em crianças, a Haemophilus influenzae, com aproximadamente 1740 genes e 500.000 aminoácidos. O fermento Saccharomyces cerevisiae com 8220 proteínas e em torno de 2.900.000 aminoácidos. O Methanococcus jannaschii, que vive nos ventos submersos muito quentes[5] com aproximadamente 450.000 aminoácidos. O genoma humano (incompleto), sendo utilizados 5733

[4] No contexto de tamanho 5 a seqüência LLLLL apareceu 37.000 vezes mais que a `WWWWW` que foram respectivamente as seqüências de maior e menor freqüências encontradas pelos autores.

[5] Provavelmente o artigo esteja se referindo aos minivulcões descobertos por Robert Ballard em 1977, que ficam em torno de 2400 m de profundidade, perto das ilhas Galápagos, e jorram uma mistura de água com ácido sulfídrico a uma temperatura em torno de $85^{o}C$.

genes humanos em torno de 3.300.000 aminoácidos.

O $\log_2 20 = 4{,}322$ deveria ser o número de bits por aminoácido, caso houvesse uma distribuição aleatória. Aplicando a técnica de blending com o método de escape PPMD e o GZIP para os genomas acima, a estimativa da entropia ficou acima de 4,322 bits por aminoácidos. Utilizando o CP para ordens -1 a 3 foi[6] possível diminuir a estimativa para valores em torno de 4,12 bits por aminoácido (exceto na ordem -1 que ficou com 4,322).

Aparentemente, para as proteínas somente é possível alcançar uma pequena taxa de compressão quando são utilizadas técnicas que necessitam da dependência de Markov. Apesar de ter sido usado um domínio do conhecimento do DNA (probabilidade de mutação dos aminoácidos) não foi possível conseguir uma boa compressibilidade para as proteínas. Uma explicação para este fato pode estar na combinação da aleatoriedade das mutações genéticas com uma representação altamente condensada de informação genética contida nas proteínas.

A técnica do compressor CP pode ser empregada para outros campos que utilizam informação, mudando-se o contexto do conhecimento do domínio (substituindo a tabela de mutação genética).

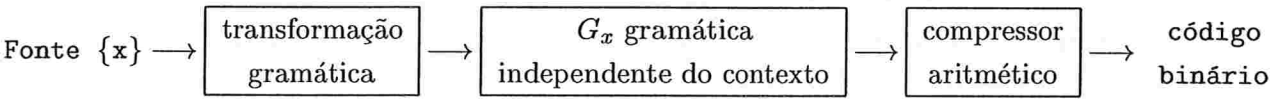
[6]Ordem -1 utiliza exatamente 4,322 bits por aminoácidos. A ordem zero mantém a independência mas sem uma distribuição uniforme. Ordem 1 com os pesos conforme a tabela de mutação (Tabela 3) etc...

CAPÍTULO 9

Entropia Utilizando Transformações Gramaticais

Uma gramática é um conjunto de regras que pode gerar todas as sentenças aceitáveis de uma linguagem. As gramáticas têm sido utilizadas como ferramentas para expressar as mensagens biológicas escritas no DNA e no RNA.

A utilização de um modelo gramático (seção 4.4) por Lanctot, Li e Yang [28] acarretou uma melhora na estimativa da entropia da estrutura primária do DNA. Foi criado o algoritmo GTAC (sigla em inglês para Compressor de Análise de Transformações Gramáticas) que efetiva uma transformação gramática na mensagem e posteriormente utiliza o resultado desta transformação em um compressor aritmético (seção 5.4) de ordem zero.



São comparados os resultados do GTAC com os obtidos através dos algoritmos Biocompress2 [25] (seção 12.3), Match Length Entropy Estimator (seção 6.2) de Farach et all [19] e com o compressor CDNA (seção 7.2) de Loewenstern e Yianilos [31].

9.1 Modelo Gramático

O GTAC conseguiu um bom desempenho porque efetiva a transformação da mensagem em uma gramática irredutível, cuja utilização será analisada, sem demonstrações.

Uma *gramática independente de contexto* (CFG) é uma quadrupla $G = \{V, \Sigma, R, S\}$ onde V representa um conjunto finito de *variáveis* ou símbolos não terminais, $S \in V$ é o *símbolo inicial*,

Σ disjunto de V e de S é um conjunto não vazio de *símbolos terminais* e R um conjunto de *regras* que associa os elementos de V a elementos de $(V \cup \Sigma)^*$.

Uma *gramática admissível* G_x é uma gramática independente de contexto que segue quatro regras básicas. A linguagem gerada por G_x consiste apenas de x. G_x é *determinístico* se qualquer variável aparece apenas uma vez à esquerda da regra que a define. O lado direito de nenhuma regra é o conjunto vazio. G_x não possui símbolos sem uso, o que só é garantido no processo de criação da variável de início S, na qual cada regra é utilizada pelo menos uma vez. Se a gramática for admissível é possível restaurar x a partir de G_x.

A *gramática irredutível* é a gramática admissível que possui as três propriedades seguintes. Cada variável v ($v \in V \mid v \neq S$) aparece do lado direito do conjunto de regras R pelo menos duas vezes[1]. Não há padrão repetido de tamanho maior que um para nenhuma regra. Cada variável distinta de G_x representa uma subcadeia distinta de x.

Um exemplo de gramática irredutível para a mensagem

$$x = \texttt{aataaatgcaatatatatgc}$$

$$\begin{aligned} S &\rightarrow BADBCCD \\ A &\rightarrow \texttt{aa} \\ B &\rightarrow A\texttt{t} \\ C &\rightarrow \texttt{at} \\ D &\rightarrow C\texttt{gc} \end{aligned}$$

Seja $w(G_x)$ a seqüência obtida pela concatenação do lado direito de todas as regras, em uma ordem qualquer, eliminando a primeira ocorrência de cada variável[2] e $|w(G_x)|$ o tamanho de $w(G_x)$. Uma possibilidade para o exemplo acima é $w(G_x) = BCD\texttt{aa}A\texttt{at}C\texttt{gc}$, e $|w(G_x)| = 12$. A entropia não normalizada $H(G_x)$ da gramática G_x é definida por

$$H(G_x) = \sum_s n(s) \log_2 \frac{|w(G_x)|}{n(s)} = -\sum_s \frac{n(s)}{|w(G_x)|}$$

onde $n(s)$ é o número de vezes que a variável ocorre em $w(G_x)$. Para o exemplo acima $H(G_x) =$ 34,26.

Dada uma gramática irredutível G_x para x, o estimador de entropia normalizado $H(G_x)/|x|$ fornece a estimativa em bits por símbolo da mensagem x.

Os principais teoremas utilizados são:

[1] A exclusão de S desta regra é importante pois o processo de conversão de x em G_x começa por $S = x$ fazendo as alterações em x à medida que as regras são criadas.

[2] Cada variável aparece pelo menos duas vezes no lado direito das regras.

- O número de bits utilizado por um compressor aritmético de ordem zero para comprimir G_x não é superior a $H(G_x) + 5|G_x| + \lambda$, onde $|G_x|$ representa a quantidade de elementos à direita de todas as regras e λ o tamanho do alfabeto da mensagem ($\lambda = 4$ para as cadeias de DNA).

- Seja x uma seqüência e $G(x)$ o conjunto de todas as gramáticas irredutíveis que representam x.

  - Há uma constante c que depende somente do tamanho do alfabeto (λ) da mensagem, que para qualquer seqüência x

    $$\max_{G_x \in G(x)} |G_x| \leq \frac{c_\lambda |x|}{\log_2 |x|}$$

    onde $|x|$ é o tamanho de x.

  - Para uma mensagem $\{X_i\}_{i=1}^{\infty}$ ergódica e estacionária com entropia H

    $$\max \left\{ \left| \frac{|B(G_{X_i})|}{n} - H \right| : G_{X_i} \in G(X_1 \ldots X_n) \right\}$$

    tende a zero com probabilidade um se $n \longrightarrow \infty$. Onde $|B(G_{x_i})|$ representa o tamanho do código binário para G_{x_i} e G_{x_i} é uma gramática admissível.

- Seja $\{X_i\}_{i=1}^{\infty}$ uma mensagem qualquer, d uma constante positiva qualquer, $X^n = X_1 \ldots X_n$ e $P(X^n)$ a probabilidade de X^n. Para qualquer gramática G_x para x há uma probabilidade de pelo menos $1 - n^{-d}$ de

  $$\frac{|H(G_{X^n})|}{n} \geq -\frac{1}{n} \log_2 P(X^n) - \frac{f(G_{X^n})}{n} - \frac{d \log_2 n}{n}$$

  onde $-\frac{1}{n} \log_2 P(X^n)$ representa a entropia da mensagem em bits por símbolo e $f(G_{X^n})$ o custo computacional (em bits) pago pela universalidade dos códigos gramáticos.

Embora achar a gramática irredutível de entropia mínima é um problema NP-difícil foi desenvolvido o GTAC que roda em tempo linear.

9.2 GTAC

O GTAC recebe como entrada uma cadeia x, constrói uma gramática irredutível G_x e comprime G_x utilizando um compressor aritmético de ordem zero.

O GTAC[3] está baseado no problema de que seja encontrado o maior padrão repetitivo não sobreposto, ou seja, na determinação das maiores subcadeias[4] iguais (mesmo tamanho e mesma

[3] O GTAC pode ser preparado para buscar o complemento reverso das seqüências do DNA, porém esta possibilidade não será detalhada.

[4] Subcadeias não vazias e não unitárias

seqüência de símbolos) e disjuntas entre si. Este problema será designado de prob1. Alguma das subcadeias solução de prob1 será chamada de *padrão*. No processo de encontrar uma gramática irredutível G_x para x prob1 aparece na busca pelo lado direito de todas as regras $r \in R$ de duas formas: se o padrão ocorre na mesma regra r_i mais de uma vez, é criada uma nova regra r_k e reescrita a regra r_i utilizando r_k em substituição ao padrão. O mesmo ocorre quando o padrão é encontrado em regras distintas, $r_i, \ldots, r_j$ uma nova regra r_k é criada e as regras $r_i, \ldots, r_j$ são reescritas utilizando r_k no lugar do padrão. O objetivo principal do GTAC é localizar cada padrão com tamanho maior que um e substituí-los por novas regras, até que somente haja padrões repetitivos de tamanho unitário. O GTAC inicialmente define a regra $S \rightarrow x$ e segue sucessivas vezes localizando os padrões, criando novas regras e substituindo todas as ocorrências dos padrões localizados pelas novas regras criadas, até que os maiores padrões repetitivos não sobrepostos (do lado direito de todas as regras) tenham exatamente tamanho 1.

As soluções triviais para prob1 exigem tempo $O(n^3)$, sendo n normalmente na ordem de milhões. A utilização de árvores de sufixo são uma saída natural para este problema, pois permite solucioná-lo em tempo amortizado $O(1)$. O GTAC[5] utiliza árvores de sufixo em sua implementação, além de outras estruturas mais simples.

9.3 Comparações e resultados do GTAC

Com o conhecimento da seqüência primária do DNA de diversos organismos, existe o desafio de reconhecer e caracterizar as regiões distintas do DNA e suas respectivas funções. A entropia é uma ferramenta que pode ajudar a resolver este problema. Para sintetizar as proteínas, as células fazem cópias temporárias de porções do DNA o mRNA, no qual cada conjunto de três símbolos representa um aminoácido. Nos seres eucarióticos, muito do código do mRNA não é utilizado na síntese protéica. As mutações aleatórias são mais degenerativas se ocorrerem nos exons ao invés de nos íntrons, (capítulo 6) havendo assim expectativas distintas para a estimativa da entropia do DNA destas regiões.

Os algoritmos LZ, Huffman e compressores aritméticos (capítulo 5) têm falhado na tentativa de obter boas taxas de compressão para a seqüência primária do DNA.

É questionado o CDNA de Loewenstern e Yianilos [31] pois é suposto que este subestima a estimativa da entropia, como no exemplo de duas seqüências aleatórias iguais rr que teriam na validação cruzada (capítulo 7) uma estimativa da entropia em torno de zero, pelo CDNA, quando de fato esta fica em torno de um. É considerado que o CDNA foi concebido como estimador de entropias mas que não há provas de sua convergência. A estimativa da entropia

[5]Não cabe neste trabalho o estudo de árvores de sufixo, de forma que a implementação do GTAC sequer será mostrada.

do GTAC ficou maior que a do CDNA, mas ficou menor na maioria dos casos à do compressor CDNA. Os algoritmos Biocompress-2 e Unix-compress superestimam o valor da entropia.

O Match Length Entropy Estimator de Farach et all [19] (capítulo 6) somente é válido se for suposto que a mensagem foi gerada por um processo de Markov (seção 3.2), sendo por isto não universal. Foram feitos alguns testes com o *E. coli* e uma estimativa da entropia de 1,85 bits/base foi obtida para os exons (4.090.525 bases) e de 1,80 bits/base para os íntrons (640.039 bases), confirmando que a entropia dos íntrons é melhor (menor) que a dos exons (Farach et all [19]).

A tabela abaixo compara a estimativa da entropia do DNA do GTAC com diversos algoritmos.

Sequence name	Sequence length	Unix compress	Bio-compress-2	CDNA compress	GTAC
PANMTPACGA	100314	2,12	1,88	1,85	1,74
MPOMTCG	186609	2,20	1,94	1,87	1,78
CHNTXX	155844	2,19	1,62	1,65	1,53
CHMPXX	121124	2,09	1,68	-	1,58
SCCHRIII	315339	2,18	1,92	1,94	1,82
HUMGHCSA	66495	2,19	1,31	0,95	1,10
HUMHBB	73308	2,20	1,88	1,77	1,73
HUMHDABCD	58864	2,21	1,88	1,67	1,70
HUMDYSTROP	38770	2,23	1,93	1,93	1,81
HUMHPRTB	56737	2,20	1,91	1,72	1,72
VACCG	191737	2.14	1,76	1,81	1,67
HEHCMVCG	229354	2,20	1,85	-	1,74

A conclusão é que a utilização de algoritmos baseados em gramáticas independentes de contexto na compressão são normalmente assintóticos, compactos e universais. O GTAC, além de obter as melhores taxas de compressão, consome tempo linear e é universal mesmo sem supor que a mensagem seja ergódica.

Capítulo 10

Compressão e Árvores Filogenéticas

Chen, Kwong e Li [10] apresentam o algoritmo GenCompress que não só consegue uma boa taxa de compressão como também uma utilização prática na comparação de genomas e na construção de árvores de evolução. É feita uma comparação do GenCompress com o Biocompress-2 e com o Cfact na qual o GenCompress melhora significativamente a estimativa da entropia das seqüências do DNA.

10.1 Biocompress-2 e Cfact

Um compressor é uma ferramenta para detectar padrões de repetição aproximados em seqüências de DNA. Antes de apresentar as base do GenCompress foi colocado o esquema de funcionamento dos compressores LZ (seção 5.5), do Biocompress-2 e do Cfact.

Os algoritmos LZ são baseados em comparações exatas e são universais. De forma genérica, eles podem ser vistos como programas com dois componentes: o primeiro faz o fraseamento da entrada em função das entradas existentes no dicionário e o segundo substitui o prefixo do que está sendo codificado por sua representação binária adequada. A concatenação destas representações binárias geram a saída do algoritmo.

O Biocompress-2 foi desenvolvido por Grumbach e Tahi [25]. Utilizam a estrutura dos compressores LZ. Busca por padrões repetitivos e seqüências repetidas invertidas (seção 2.6) já examinadas e codifica o tamanho e a posição da ocorrência. Caso não haja uma repetição significativa, o Biocompress-2 utiliza um compressor aritmético de ordem 2 para codificar o próximo símbolo.

O Cfact foi criado por É. Rivals, et al. [45] Procuram as maiores cadeias repetidas, utilizando a estrutura de dados chamada árvore de sufixo em toda a seqüência. É um algoritmo dividido em dois passos. No primeiro é criada a árvore de sufixo. No segundo, a codificação, caso haja

um ganho efetivo na representação das seqüências repetidas, estas são representadas através do método LZ; caso contrário são utilizados dois bits por símbolo na codificação.

Os processos envolvendo o Biocompress-2 e o Cfact são similares. A principal diferença está no fato de o Biocompress-2 utilizar um compressor aritmético de ordem 2 e o Cfact não.

10.2 GenCompress

O GenCompress faz comparações entre seqüências e, se estas forem similares, ele pode utilizar transformações para torná-las iguais. São três tipos de operações utilizadas nas transformações. A substituição do símbolo na posição p pelo símbolo c, a inserção do o símbolo c na posição p e a exclusão do símbolo que ocupa a posição p. Serão representadas: a substituição pela letra S, a inserção pela letra I e a exclusão pela letra E. Há pelo menos duas formas de converter a seqüência `gaccttca` em `gaccgtca`. A primeira forma está na substituição do elemento da posição 4 (inicia-se a cadeia na posição zero) por g, a qual pode ser representada por $\{(S, 4, \mathrm{g})\}$. A segunda forma está na inserção do elemento g na posição 4 e exclusão do elemento na posição 6, que pode ser representada por $\{(I, 4, \mathrm{g}), (E, 6)\}$.

São analisadas algumas formas de representação da seqüência `gaccgtca` a partir da `gaccttca`. A primeira forma de representação utiliza 2 bits por símbolo (independe da seqüência `gaccttca`). Por exemplo na utilização 00 para representar `a`, 01 para `c`, 10 para `g` e 11 para `t`, a seqüência é representada por

$$\{10\ 00\ 01\ 01\ 10\ 11\ 01\ 00\}$$

utilizando deste modo 16 bits em sua representação (os espaços foram adicionados para facilitar a visualização). Uma segunda forma de representação através da comparação exata, indica a posição inicial e o tamanho da subcadeia repetida. Desta forma é colocada a seqüência `gaccttca` nas posições zero a sete da memória. Utiliza-se três bits para representar um inteiro, dois bits para representar um algarismo (conforme a forma anterior de representação) e um bit para representar se há uma repetição (0 indica um símbolo e 1 indica uma referência à memória). A cadeia `gaccgtca` é representada por $\{(0, 4), \mathrm{g}, (5, 3)\}$ que em bits assume

$$\{1\ 000\ 100\ 0\ 10\ 1\ 101\ 011\}$$

ocupando um total de 17 bits. A terceira forma de representação da cadeia utiliza a comparação inexata, levando em conta a primeira transformação utilizada $\{(S, 4, \mathrm{g})\}$ e a cadeia `gaccttca` nas posições 0 a 7 da memória. Neste caso S é representado por 00, I por 01 e E por 11. Um bit é necessário para indicar se a próxima operação é uma referência à memória (0) ou uma operação de substituição (1). A seqüência `gaccgtca` é representada por $\{(0, 8), (S, 4, \mathrm{g})\}$ que gera em binários

$$\{0\ 000\ 111\ 1\ 00\ 100\ 10\}$$

utilizando 15 bits. A última forma de representação é parecida com a terceira, mas leva em consideração as operações $\{(I, 4, \mathrm{g}), (E, 6)\}$ na conversão das cadeias. Desta forma a representação é $\{(0, 8), (I, 4, \mathrm{g}), (E, 6)\}$ que em binário fica

$$\{0\ 000\ 111\ 1\ 01\ 100\ 10\ 1\ 10\ 110\}$$

consumindo um total de 21 bits. Para este exemplo a forma de representação mais curta utilizou a comparação inexata.

O GenCompress se baseia em comparações inexatas e procede da seguinte forma. Considera a mensagem dividida em dois seguimentos: o primeiro (inicialmente vazio) contém o que já foi codificado e o segundo (vazio ao final da execução do algoritmo) contém o que falta ser codificado. O *prefixo ótimo* é o segmento que ocupa os primeiro bytes (prefixo) do segundo segmento, cuja codificação (aplicando as operações de edição, se for o caso) utiliza o menor número de bits (é ótima). Este programa localiza o prefixo ótimo do segundo segmento que é aproximadamente igual a alguma subcadeia do primeiro segmento e deste modo é codificado de forma econômica. Após sua codificação este prefixo ótimo é retirado do segundo segmento e concatenado ao final do primeiro. A operação se repete até que o segundo segmento esteja vazio. Na codificação das regiões nas quais o prefixo ótimo gera mais de 2 bits por símbolo é utilizado um compressor aritmético de ordem 2. Nas demais regiões é utilizada uma representação apropriada de inteiros positivos e suas operações de edição.

Embora para o GenCompress o desempenho e a complexidade estejam em segundo plano, o tamanho das cadeias de DNA acabam por obrigar a utilização de otimizadores na busca do prefixo ótimo. Uma das otimizações mostra que o prefixo ótimo sempre termina antes de um símbolo que não coincide. Para a seqüência base `aaacccgtgt` (caso haja apenas uma cadeia base) e para a cadeia a ser codificada `acacctgtaa`, o prefixo ótimo está entre uma das três opções `a`, `acacc`, `acacctgt`. Desta forma a busca do prefixo ótimo no GenCompress começa pela busca através de uma comparação exata com tamanho pequeno prefixado e a partir daí passa a buscar as comparações inexatas. A estrutura de dados utilizada é a árvore de sufixo.

Dentro desta proposta foram criadas duas variações do algoritmo: uma que utiliza somente operações de substituição o GenCompress-1 e outra que utiliza as três operações o GenCompress-2. Ambas obtiveram resultados similares na estimativa da entropia das seqüências do DNA. Na comparação com o Biocompress-2 e o Cfact houve ganhos significativos na diminuição da estimativa da entropia.

10.3 Compressores e árvores filogenéticas

Árvores filogenéticas são as representações através das árvores da história evolucionária das espécies. Em muitos métodos de construção de árvores filogenéticas, o primeiro passo é definir

uma função distância para determinar a proximidade entre as cadeias de DNA. Parte-se da hipótese de que seqüências que estão próximas com relação à distância adotada também devem estar próximas na árvore de evolução. Existe uma grande dificuldade em definir "distâncias" entre seqüências de DNA. Um problema é que as distâncias não são simétricas. Os pesquisadores têm buscado encontrar uma medida de distâncias entre seqüências de DNA que seja apropriada à construção de árvores filogenéticas. Uma alternativa natural é a

$$\text{compressão}(u|v),$$

na qual o compressor é calibrado na seqüência v e utilizado na seqüência u. Calibrar o compressor significa utilizar uma seqüência para gerar uma distribuição de probabilidades ou um dicionário. Infelizmente compressão$(u|v)$ geralmente é diferente da compressão$(v|u)$, ou seja, não é simétrica.

Chen, Kwong e Li [10] utilizaram o teorema da complexidade de Kolmogorov e chegaram à seguinte função distância:

$$\text{dist}(u, v) = \text{compressão}(u) - \text{compressão}(u|v) \tag{10.1}$$

esta função é "aproximadamente" simétrica e é uma aproximação da informação mutual (seção 3.4).

O GenCompress foi adaptado para realizar alguns experimentos através de uma aplicação da função dist$(u|v)$ para estimar o grau de correlação entre as seqüências u e v. Os resultados preliminares são encorajadores. Notar a semelhança entre a expressão (10.1) e a entropia mutual (equação (3.3)).

Capítulo 11

Aplicações Biológicas de Estimadores de Entropia

Além da utilização da estimativa da entropia para fins de compressão, alguns trabalhos utilizaram os algoritmos nas seqüências do DNA para outras finalidades. Os objetivos variaram muito de artigo para artigo de forma que não foi possível fazer comparações entre os diversos algoritmos. A opção foi fazer uma análise da aplicação e do algoritmo (quando possível).

11.1 Match Length Entropy Estimator

Este algoritmo desenvolvido por Farach et al. [19] (capítulo 6) se baseia na metodologia dos compressores LZ (seção 5.5). Ele foi utilizado para a estimativa da entropia de genes humanos e na tentativa de resolver o problema da detecção do limite íntron/exon. A mutação genética é aleatória e tende a ocorrer com a mesma freqüência nos íntrons e nos exons. *A priori* se pensava que os íntrons, por serem menos sujeitos às pressões da seleção natural, seriam mais sensíveis à aleatoriedade das mutações genéticas que os exons. Previa-se uma menor estimativa da entropia para os exons que para os íntrons. Mas as medidas feitas confirmaram exatamente o contrário: os íntrons são mais previsíveis que os exons.

Dois dos aspectos do DNA foram destacados: a necessidade de rápida convergência do estimador de entropia, pois no DNA as cadeias de íntrons e exons são relativamente curtas[1] (em especial a dos exons) e previsibilidade dos íntrons. Concluiu-se que tanto os íntrons quanto os exons estão sujeitos às forças restauradoras e que parte dos íntrons poderiam servir para definir e determinar o limite íntron/exon. As regras de codificação envolvidas neste processo são com-

[1] Para os genes humanos estudados, o tamanho médio dos íntrons foi 867 com desvio padrão de 583 e o tamanho médio dos exons foi de 184 com desvio padrão de 96. Porém as medianas foram muito baixas, sendo para os íntrons 434 e para os exons 139.

plexas e exigiriam uma base de testes muito maior para sua determinação. Farach et al. [19] concluíram citando as forças restauradoras mas omitiram qualquer explicação sobre elas.

Uma análise superficial sobre a entropia dos íntrons e dos exons induziria à hipótese de que, como os íntrons são maiores que os exons, eles seriam mais previsíveis. Se esta fosse confirmada, na utilização de uma grande quantidade de genes de uma mesma espécie ocorreria uma diminuição da estimativa da entropia dos exons. Mas Nevill-Manning e Witten [39] tentaram sem sucesso comprimir as proteínas de algumas espécies. Foram utilizados 5.733 genes humanos (além de genomas completos de três espécies) nesta tentativa, mas a estimativa da entropia dos exons não foi reduzida. Não se conhece uma explicação biológica para este fato ainda. Uma hipótese possível é que o processo de seleção natural "otimizou" os códigos genéticos das proteínas (exons). Para este organismo se manter, ele gasta menos energia no processo da síntese protéica que os concorrentes da mesma espécie. Como as proteínas representam mais da metade do peso seco de uma célula, a eliminação de redundâncias pode ter sido um fator de vantagem no processo da adaptação ao meio e ter sido relevante na evolução[2] das espécies.

O Match Length Entropy Estimator também foi utilizado na detecção do limite íntron/exon. Os íntrons quase sempre começam com gt e terminam com ag. Foram detectados quatro padrões para o início do íntron:[3] aaggtaagt, aacgtgagt, gacgtaagt e caggtgagt. Para estes a entropia ficava abaixo de 1,8 e em outras regiões próximas aos pares gt a entropia ficava entre 1,8 e 2,0. Porém o algoritmo não obteve sucesso para detectar padrões de final de íntron.

Aparentemente a falha na detecção de final de íntron contradiz a hipótese proposta por Farach et al. [19] da existência de uma regra nos íntrons para definir e determinar o limite íntron/exon. Caso esta se confirme a questão deste limite ficará definitivamente resolvida e ao íntron caberá a determinação dos segmentos codificáveis do DNA. No atual desenvolvimento dos estudos não se pode atribuir esta função aos íntrons, cuja finalidade continua desconhecida.

11.2 Os transportadores de informação

Ebeling e Frömmel [17] desenvolveram um trabalho sobre três transportadores de informação distintos: cadeias de DNA e proteínas, obras literárias e músicas clássicas. O trabalho deles foi enfocado na informação mutual (transinformação) e nas correlações de longas distâncias.

Para a análise das seqüências de DNA, os genomas de diversos seres foram estudados. O algoritmo adotado utilizou o método LZ mas não teve seus detalhes publicados neste artigo [17],

[2] Não havendo redundância, gasta-se menos energia e substâncias na síntese protéica, além de minimizar a probabilidade de erro de codificação. O organismo, ao utilizar menos energia, necessita de menor quantidade de alimentos e com isto consegue sobreviver em ambientes com escassez de comidas.

[3] A representação do gt sublinhado indica o ponto chave da comparação, ou seja estas duas bases indicam o início do íntron.

o que impossibilitou a análise do algoritmo utilizado. Este também foi utilizado nos textos Moby Dick, Contos de Grimm e na Bíblia. Tanto nas cadeias do DNA quanto nos textos adotados houve correlações de longa distância com ondas de flutuação de tamanho entre 102 e 103. A homogeneização dos resultados impede observação das flutuações.

Na análise das cadeias de proteínas, o alfabeto teve seu tamanho ajustado de 4 para 20. Foi constatado que alguns aminoácidos ocorrem com maior freqüência que outros e foram encontrados picos nos quais a previsibilidade da cadeia era maior que fora destes.

Já nas cadeias musicais houve uma diferença significativa entre os diversos autores das músicas analisadas. Porém não foi possível determinar se a música de Mozart era mais ou menos organizada que a de Beethoven.

A conclusão foi que existem correlações de diversas distâncias nas cadeias de DNA, inclusive as correlações de longas distâncias.

11.3 Árvores filogenéticas

As árvores filogenéticas têm sido uma das aplicações dos algoritmos computacionais na biologia. Chen, Kwong e Li [10] ao criarem o GenCompress mostraram a possibilidade da utilização deste programa para este fim. Leitner et al. [29] fizeram a reconstrução da história de transmissão conhecida do vírus HIV-1 através da análise da árvore filogenética.

GenCompress

No processo de construção das árvores filogenéticas, o primeiro passo é calcular a distância entre os pares de cadeias de DNA. O GenCompress utiliza o teorema da complexidade de Kolmogorov para calcular a informação mutual. Com isto se consegue uma distância simétrica entre os genomas.

O GenCompress foi aplicado em algumas cadeias genéticas e a partir dos dados obtidos foram feitas tabelas de correlações e construídas duas árvores filogenéticas. A entropia mutual (seção 3.4) possui uma expressão bem próxima à utilizada na determinação da distância entre as cadeias genômicas. Há um indicativo que ela pode auxiliar na construção desta árvores. Os resultados preliminares são promissores, porém Chen, Kwong e Li [10] não foram conclusivos quanto ao emprego do GenCompress no auxílio da construção de árvores filogenéticas, ficando em aberto sua utilização prática para este fim.

Aplicações no HIV-1

Leitner et al. [29] aplicaram a análise das árvores filogenéticas nos vírus HIV-1 para reconstruir a história de sua transmissão. Eles conseguiram chegar a resultados muito próximos da história real da transmissão do vírus utilizando várias técnicas de construção de árvores filogenéticas, mas não citaram a entropia. A conclusão dos autores foi que a utilização de árvores filogenéticas podem ser utilizadas para auxiliar na resposta a questões complexas a respeito do HIV-1.

Embora a construção de árvores filogenéticas não esteja diretamente ligada à estimativa da entropia Chen et al. [10] quando criaram o GenCompress pensaram em uma dupla possibilidade de utilização: compressão e construção de árvores filogenéticas. A informação mutual serviu de elo entre as duas utilizações. A busca da distância simétrica entre os genomas e a estrutura de comparação inexata deste algoritmo possuem alguns pontos em comum.

11.4 Regiões de baixa entropia nas cadeias de DNA

Crochemore e Vérin [13] criaram um método para detectar regiões do DNA com repetições incomuns. Definiram janelas com tamanho de 500 nas posições múltiplas de 250. Houve desta forma sobreposição de 50% de uma janela em relação à sua sucessora. Para restringir a busca de padrões foram definidas cadeias com tamanhos maiores que 8. O algoritmo foi utilizado nos cromossomos do *Saccharomyces cerevisiae*. As medições estruturadas desta forma permitiram uma análise rápida e eficiente das cadeias do DNA, capturando as pequenas repetições tanto dos padrões longos quanto dos curtos. Desta forma foi definido um perfil associado a cada cromossomo do fermento.

O método utilizado é um primeiro passo para o desenvolvimento de medições mais precisas nas cadeias de DNA e se mostrou útil tanto na determinação de novas características biológicas do *Saccharomyces cerevisiae* quanto em outras classes de seqüências de DNA.

Os estimadores de entropia podem ser utilizados para identificar as características biológicas do DNA. Sua aplicabilidade depende da necessidade dos biólogos. O trabalho em conjunto dos cientistas do DNA com os matemáticos e cientistas da computação permite ajustar o conhecimento biológico à tecnologia da informação. Este gera um desenvolvimento aplicado da teoria da informação às necessidades práticas da biologia do DNA.

Capítulo 12

Compressores para Seqüências de DNA

A maioria dos artigos analisados neste texto criou algoritmos para trabalhar nas seqüências do DNA. Alguns documentos não descreveram o algoritmo utilizado tornando inviável, até certo ponto, a análise do modelo adotado. Cada trabalho destinou seus programas a um fim específico, mas foi predominante a busca da diminuição da estimativa da entropia, cada qual com suas otimizações e resultados. O objetivo deste capítulo é detalhar como cada algoritmo tratou o DNA e como este "reagiu" ao tratamento dado. A primeira análise será sobre os compressores convencionais, passando pelas melhorias obtidas por cada algoritmo e seus respectivos ganhos. As aplicações dos programas variaram de acordo com o algoritmo. Nem sempre é possível tabular os resultados, mas no geral estes podem ser enfocados.

Embora o ênfase deste capítulo não seja a estrutura de dados, cabe um destaque para a estrutura particular chamada "árvore de sufixo". Esta tem sido utilizada nas cadeias genômicas, como nos artigos [28, 10], nos escritos citados por Gusfield [26, capítulos 5 a 9] e no algoritmo Cfact descrito por Chen et al. [10]. O DNA tem "respondido" bem a esta estrutura, pois os algoritmos que a utilizaram melhoraram a estimativa da entropia dentro do escopo por eles abrangido.

Loewenstern e Yianilos [31] comentaram a respeito da estimativa de entropia do DNA. Com o desconhecimento do processo de "geração" da mensagem (o DNA), a estimativa da entropia está sendo efetivada a partir da saída do compressor, computando assim a quantidade de bits/nucleotídeos obtidos pelo algoritmo. Eles (Loewenstern e Yianilos) fizeram algumas considerações a respeito desta forma de estimar a entropia do DNA, com ênfase no custo adicional que isto representa e propondo alternativas métricas para este processo.

Os compressores seguem dois paradigmas principais em seu funcionamento: as técnicas probabilísticas e dicionários (capítulo 4). As técnicas probabilísticas, baseadas nas distribuições

de probabilidades, determinam a previsibilidade do próximo símbolo a partir da análise das situações passadas (estado ou contexto). Por exemplo, se os três últimos símbolos codificados foram aac, o processo busca todos os sucessores de aac que já ocorreram (distribuição de probabilidades ou contadores) gerando desta forma a previsibilidade do próximo símbolo. Em favor desta técnica está o fato de que nas seqüências do DNA ocorrem mais repetições próximas do que é esperado em cadeias aleatórias[1]. Há dois fatos que dificultam a utilização das técnicas probabilísticas: mutações genéticas são aleatórias e quanto maior for o contexto, muito menor é a possibilidade dele haver ocorrido[2]. Nas técnicas de dicionário, cada frase é representada por uma entrada no dicionário, que normalmente ocupa menos espaço que a própria frase. Quanto maior contexto muito menor a possibilidade deste ter aparecido anteriormente; este fato diminui a eficiência deste paradigma, pois os contextos maiores são os melhores candidatos para as entradas nos dicionários (quanto maior a frase maior o ganho na compressão). Lanctot et al. [28] mencionam que algoritmos tradicionais como código de Huffman e vários outros baseados no método LZ falharam na tentativa de comprimir as seqüências do DNA.

12.1 Algoritmos probabilísticos

Os algoritmos probabilísticos geralmente seguem a distribuição da probabilidade do processo de Markov [39] (seção 3.2). A compressão aritmética é o algoritmo baseado nos modelos probabilísticos mais aplicado às cadeias de DNA. É comum a utilização de processos adaptativos ou semi-adaptativos no tratamento da probabilidade do próximo símbolo da seqüência genômica. Os modelos empregados nos compressores para as cadeias polinucleotídicas normalmente adotam alguma ordem do modelo de contextos finitos (seção 4.2) ou alguma das técnicas de blending (seção 4.6).

O DNA possui informação. Teoricamente deveria utilizar menos de dois bits por símbolo para representá-lo. Aplicado no Humretblas[3] um modelo de contextos finitos de ordem zero (os símbolos do alfabeto não ocorrem de forma homogênea no gene) obteve a estimativa da entropia em torno de 1,95 bits por caractere, bem próximo da distribuição aleatória. Para as linguagens naturais, ao se aumentar a ordem do modelo, se reduz consideravelmente a estimativa da entropia da mensagem. Porém no caso do Humretblas a melhor estimativa de entropia ficou em 1,90 bits por caractere (ainda bem próximo da distribuição aleatória), conforme artigo de Mantegna et al. [34] (citado por Loewenstern e Yianilos [31]). Os modelos de contextos finitos não ajudam muito na melhora (diminuição) da estimativa da entropia do DNA.

---

[1] Loewenstern e Yianilos [31].

[2] Problema da freqüência zero (seção 4.1).

[3] Gene humano da susceptibilidade ao retinoblastoma com 180.388 nucleotídeos [31].

CDNA

Loewenstern e Yianilos [31] desenvolveram dois algoritmos parecidos para a estimativa da entropia do DNA. O primeiro deles, o CDNA, serve somente como estimador de entropia. O segundo deles é o compressor CDNA, que utiliza o modelo do CDNA e gera uma compressão para as cadeias genômicas. O CDNA trabalha de forma semi-adaptativa (cria as distribuições de probabilidades para a mensagem inteira antes de calcular a estimativa da entropia). O compressor CDNA trabalha de forma adaptativa (vai ajustando as probabilidades enquanto a mensagem vai sendo codificada). Assim foi possível, a partir do aprendizado obtido, fazer a compressão. A estimativa da entropia das cadeias genômicas feitas pelo CDNA sempre foram melhores (menores) que as obtidas pelo compressor CDNA. Isto indica que sempre há perdas no processo de compressão, porém as perdas foram pequenas na maioria dos casos (somente em 2 casos foi significativa).

O CDNA é baseado na comparação inexata (distância de Hamming), a partir da qual é possível fazer distribuições de probabilidades de ordens maiores sem a preocupação com o problema da freqüência zero. Quando o contexto ainda não apareceu, a estimativa das probabilidades do próximo símbolo é feita com a utilização da distância de Hamming. O algoritmo utiliza pesos na definição das probabilidades do próximo símbolo, em função da quantidade de erros na comparação[4]. Os detalhes podem ser vistos no capítulo 7. Foi comentado que em geral a compressão, utilizando o modelo de ordem 4, superou[5] a do modelo de ordem 6. Isto indica que as cadeias maiores tendem a ocorrer com pouca freqüência no DNA.

Lanctot et al. [28] questionam o CDNA. A validação cruzada[6] é obtida através da divisão da seqüência em n segmentos de tamanho igual. Cada um dos n segmentos ocupa a última posição da cadeia e é feito o treinamento nos $n-1$ segmentos restantes. A entropia é estimada pela média das n estimativas individuais. O questionamento é feito através de um contra exemplo: supondo $n = 2$ e a seqüência `rr` como a repetição da seqüência aleatória `r`. Pelo processo convencional, a entropia ficaria em torno de 1 bit/símbolo (metade da seqüência aleatória com 4 símbolos no alfabeto) e pela validação cruzada ficaria próxima a zero. É claro que este contra-exemplo mostra a fragilidade da validação cruzada para cadeias que apresentem certas características como cadeias com n segmentos iguais que se repetem sucessivas vezes. Mas este não é o caso do DNA. Mesmo que fosse, a utilização de um valor maior para n torna a deturpação do contra exemplo bem menor (embora ainda exista). Se existe algum modelo gerador das mensagens da vida (DNA), ele ainda não foi descoberto. Os compressores geram uma superestimativa da entropia pois possuem um custo que é passado para a mensagem comprimida, mas que é menor em sua seqüência final. A validação cruzada serve para homogeneizar este custo mínimo,

[4] Utiliza também a técnica "*Expectation Maximization*" de Baum-Welch na definição dos pesos.
[5] Problema da freqüência zero.
[6] *Cross validation* no original do artigo.

evitando desta forma o erro grosseiro de se tomar a última cadeia como representativa do todo. Se a última cadeia possuir uma entropia menor que o restante, a estimativa da entropia do todo estaria subestimada. Se houver uma quebra de contexto para este último segmento, a entropia estaria superestimada. A média, mesmo não sendo um valor matemático rigoroso, trará uma aproximação bastante real do valor da estimativa da entropia do todo (exceto para alguns casos anômalos).

O fato do CDNA obter a melhor estimativa de entropia para as cadeias de DNA poderia indicar de que o DNA possuiu anteriormente uma dependência maior ao processo de Markov. Com o decorrer do tempo (e mutações genéticas aleatórias) esta dependência diminuiu. Mas o fato é que as mutações genéticas devem não só ser viáveis mas representarem vantagens para serem incorporadas ao gene da nova espécie. Se representam um ganho (para a espécie no hábitat em que ela vive) é equivocada a cogitação de uma ordem anterior "perturbada" pelas alterações genômicas.

CP

Enquanto o CDNA trabalha com comparações inexatas para cadeias de DNA, o CP utiliza blending com o chamado método de escape PPMD, mas somente para os exons, com o alfabeto de 20 aminoácidos. A base probabilística do CP leva em conta as probabilidades de mutação genética de um aminoácido se transformar em outro no período de 62 milhões de anos. Para a definição dos pesos dos contextos as probabilidades de mutação genéticas são consideradas, tanto o que está fora do contexto atual quanto o conteúdo deste.

O CP faz a generalização do blending no sentido de verificar todas as ocorrências na variação. Na suposição de que o contexto `ILW` ainda não ocorreu, na utilização convencional das técnicas de blending o contexto é reduzido para `LW`, mas, baseado na tabela de mutações genéticas, o CP verifica as probabilidades dos contextos `*LW`, `I*W` e `IL*`, onde `*` representa um símbolo qualquer, antes de reduzir a ordem do contexto.

Este programa foi utilizado em proteínas de três genomas completos e em uma parte das proteínas humanas. Houve uma opção por este tipo de base para o teste do programa, pois para seres diferentes podem existir proteínas similares provocando vícios na estimativa da entropia. Assim as bases de teste do algoritmo foram sempre proteínas distintas de uma mesma espécie.

Porém, este algoritmo não conseguiu uma diminuição significativa na estimativa da entropia. Isto mostra que, de alguma forma, o fato da mutação genética ser aleatória dificulta a estimativa da entropia nas proteínas. Se as mutações genéticas tendem a ocorrer com a mesma freqüência nos íntrons e exons e as mutações nos exons tendem a sofrer mais a pressão da seleção natural que a nos íntrons [19] era de se supor que a estimativa de entropia nos exons fosse menor que nos íntrons, mas ocorre exatamente o contrário [19].

A utilização da tabela de mutações genéticas nos pesos dos contextos pode haver influído negativamente na estimativa da entropia, pois há algum fator desconhecido no processo da mutação genética em relação à informação dos genes; este fator pode ser a seleção natural. A tabela de mutações genéticas reflete em média as probabilidades de mutação dos aminoácidos. Cada mutação, tomada individualmente, pode contrariar estas probabilidades. É muito pouco provável que os aminoácidos WWW tenham sido LLL um dia, mas isto pode ter acontecido. Embora a tabela reflita a média das mutações, em vários pontos das proteínas podem ter ocorrido mutações improváveis, descaracterizando a previsibilidade destas proteínas como um todo. O tamanho das proteínas é muito pequeno para que as distorções pontuais sejam corrigidas e as distribuições de probabilidades das proteínas reflitam as probabilidades de mutação dos aminoácidos. Dado que a mutação genética ocorreu, somente os indivíduos mais adaptados WWW passaram seus genes para frente, gerando a nova espécie. A aleatoriedade da mutação genética não implica sua passagem para a nova espécie; somente esta ocorrerá se houver ganhos efetivos nos indivíduos mutantes (seleção natural). É fato que a entropia nos exons é maior que nos íntrons e sua causa é desconhecida. Talvez haja uma relação entre os exons e os íntrons que os antecedem, talvez não, mas é certo que a célula sabe reconhecer com exatidão cada íntron e cada exon e monta corretamente as proteínas dos seres vivos.

12.2 Algoritmos baseados nos compressores LZ

O compressor do Unix, uma das aplicações baseadas nos algoritmos LZ, teve um desempenho pior para as seqüências de DNA que a distribuição aleatória, gastando em média mais de dois bits por nucleotídeo. Nos artigos analisados a pior taxa de compressão coube a estes tipos de compressores. Porém a mistura de algum método LZ com outros modelos gerou bons resultados (para o DNA), embora não os melhores na compressão das cadeias do DNA.

Dos algoritmos que apresentados, três deles não foram diretamente transcritos dos artigos que originalmente os definiram: o Biocompress, Biocompress-2 e o Cfact, uma explicação rápida deles foi passada por Chen et al. [10] e as considerações sobre estes programas estarão restritas às contidas neste artigo [10].

Dois dos artigos estudados [19, 17] utilizaram programas baseados nos algoritmos LZ, mas não mostraram o resultado da estimativa da entropia do DNA obtidos destes programas. Desta forma seus algoritmos não serão tratados neste capítulo.

Biocompress

O Biocompress foi desenvolvido por Grumbach e Tahi [24] para localizar repetições exatas ou seqüências com repetições invertidas (seção 2.6) nas cadeias de DNA. Apesar de alguns

artigos [28, 31, 10] citarem as estatísticas de compressão do Biocompress-2 nenhum destes citou os dados de compressão do Biocompress, ficando os comentários sobre este programa restritos a este parágrafo.

Cfact

O algoritmo Cfact de Rivals et al. [45] trabalha de forma similar ao Biocompress, exceto que ele utiliza a estrutura de dados árvore de sufixo na seqüência inteira. É um algoritmo em duas etapas: na primeira constrói a árvore de sufixo e na segunda a utiliza garantindo que o segmento da mensagem que está sendo codificado obterá ganho na codificação; caso contrário ele será representado com dois bits por base. Como as seqüências tratadas por este algoritmo foram distintas das tratadas pelos demais, as comparações somente puderam ser feitas com o algoritmo GenCompress de Chen et al. [10] ficando neste caso em uma tabela a parte.

12.3 Algoritmos mistos

Alguns algoritmos, para melhorar sua taxa de compressão, utilizaram mais de uma técnica em seu modelo. Embora o Cfact pudesse estar nesta seção, a rigor a mistura por ele utilizada foi a de representar as subcadeias com dois bits por caractere. Isto foi considerado, para fins deste trabalho, como uma forma alternativa de representação. Se esta técnica fosse utilizada em toda a cadeia não haveria compressão nem perdas, somente uma outra forma de codificar. Por isto ele foi mantido na seção anterior.

Biocompress-2

O Biocompress-2 de Grumbach e Tahi [25] trabalha de forma similar ao Biocompress, mas utiliza um compressor aritmético de ordem 2 caso o tamanho da seqüência repetida encontrada não seja significativo. Como seus dados de compressão foram citados por diversos artigos nota-se que, apesar da utilização de uma mistura de metodologias (compressor probabilístico e compressor por dicionário), seu ganho foi relativamente pequeno em relação ao compressor aritmético convencional de ordem 4.

GTAC

Lanctot et al. [28] utilizaram transformações gramaticais na seqüência do DNA e aplicaram um compressor aritmético de ordem zero na cadeia transformada. O GTAC utiliza como estrutura principal a árvore de sufixo. Foi o compressor que conseguiu as melhores taxas de compressão

(seção 3.3) para seqüências do DNA, embora o CDNA obteve estimativas de entropia melhores que o GTAC.

O princípio básico da transformação gramatical está na modificação da seqüência para que nesta não ocorram subcadeias repetidas maiores que um. Desta forma, o compressor aritmético de ordem zero é eficiente, pois todas as ocorrências para a ordem um não ocorrem ou ocorrem somente uma vez. Desta forma haveria uma distribuição uniforme de probabilidades dos símbolos que apareceram (o que comprometeria a estimativa da entropia).

As transformações gramaticais foram totalmente reversíveis. A gramática utilizada em cada transformação (de alguma cadeia de DNA) era independente do contexto, admissível e irredutível. Quando se aplicou a gramática irredutível, todas as cadeias com tamanho maior que 1 passaram a ocorrer somente uma vez na seqüência transformada. Há dois custos principais no GTAC, a saber: a necessidade da representação da gramática junto com a mensagem original e o aumento do tamanho do alfabeto devido a transformação gramatical (o que é ruim para o compressor aritmético).

Apesar destes custos, o GTAC obteve taxas de compressão para a maioria das cadeias examinadas, melhores que o compressor CDNA. Como o CDNA foi um pouco melhor que o GTAC e no GTAC há algumas perdas no compressor, é provável que a melhor estimativa da entropia do DNA esteja no CDNA. Apesar de todos estes indicativos, o número de genes/espécies testados foi consideravelmente pequeno para que estes dados sejam considerados significativos como um todo; há somente uma indicação a este respeito.

É possível fazer uma análise sobre o bom desempenho do GTAC, mesmo com a falta de representatividade da amostra utilizada nos testes deste algoritmo: o GTAC conseguiu capturar todas as repetições (mais freqüentes que na distribuição aleatória) nas cadeias de DNA. E o fez tornando unitária qualquer subseqüência de tamanho maior que um. Desta forma, ele transformou a mensagem em outra onde não ocorrem repetições de cadeias mas somente de símbolos. Pagou o preço do aumento considerável do número de símbolos do alfabeto, mas ganhou a eficiência da captura de todas as pequenas repetições.

GenCompress

Este é um outro algoritmo que explora as árvores de sufixo e o método LZ. A implementação feita por Chen et al. [10] criou duas variantes do GenCompress: o GenCompress-1 e o GenCompress-2. Ambos são baseados em operações de edição. São três possíveis operações de edição: substituição de uma base, inclusão de uma base e exclusão de uma base. As duas primeiras necessitam de uma posição e uma base (para inclusão ou substituição); a terceira somente da posição. Estas operações são feitas sobre a árvore de sufixo que vai sendo construída à medida que é feita a codificação da cadeia de DNA. Trata-se de um algoritmo misto baseado em busca

aproximada. Se as operações de edição para a metodologia LZ não apresentarem um ganho aceitável, é feita a codificação da subcadeia com a utilização de um compressor aritmético de ordem 2. Se a busca (exata ou aproximada) atender a um ganho mínimo, a subcadeia é codificada (juntamente com suas operações de edição, se estas existirem) no método LZ.

A diferença entre o GenCompess-1 e o GenCompress-2 é que o primeiro utiliza somente as operações de substituição (distância de Hamming) e o segundo as três operações citadas.

Embora o enfoque principal do algoritmo esteja na metodologia LZ ele pode ser olhado por um outro aspecto como uma melhora no algoritmo probabilístico de ordem 2. Os métodos LZ têm sido utilizados com muito sucesso na linguagem natural (o compressor do Unix, por exemplo), embora tenham fracassado para as cadeias genômicas.

Loewenstern e Yianilos [31] forneceram um motivo para a utilização de uma ordem tão baixa no compressor aritmético para as cadeias de DNA: a taxa de compressão para o algoritmo de ordem 4 é praticamente idêntica à obtida pela ordem 6. Uma parte da cadeia está sendo codificada pela tecnologia LZ. O Biocompress-2 obteve resultados aceitáveis usando um compressor desta ordem; provavelmente ela é ótima nestas condições. O artigo [10] não mencionou a existência de estudos ou testes para definir a ordem do compressor aritmético.

12.4 Comparação dos resultados

Até agora foram descritos os comportamentos de diversos algoritmos na compressão ou estimativa da entropia das cadeias primárias do DNA e existe um resultado prático destes programas. Sua apresentação foi agrupada em três tabelas, que são edições das tabelas originais publicadas nos artigos [28, 31, 10] e organizadas para este trabalho.

A primeira traz o resultado dos algoritmos convencionais de compressão. Uma análise rápida mostra que o compressor do Unix, teve um desempenho desastroso: utilizou mais de dois bits por base, pior que a distribuição aleatória. Os compressores aritméticos apresentaram um desempenho medíocre, obtendo taxas de compressão em torno de 0,05. Os desempenhos das ordens 4 e 6 ficaram bem próximos, ficando difícil definir o melhor dentre estes.

Foi adicionada uma linha ao final de cada tabela com a média (em bits por base) dos valores obtidos, desprezando os valores para os quais o algoritmo não foi testado. Esta média serve apenas como um valor de referência, não devendo ser tomada como valor significativo do algoritmo, pois além do tamanho da amostra não significativo há contextos utilizados por cada artigo que dificultam a generalização dos resultados obtidos. Por exemplo, Lanctot et al. [10] na apresentação do GTAC fazem uma crítica sobre o CDNA. Várias medições sobre o CDNA

Cadeia	Tamanho	Compressor Unix	Ordem 1	Ordem 2	Ordem 4	Ordem 6
Humdystrop	38.770	2,23	1,95	1,92	1,91	1,95
Humghcsa	66.495	2,19	2,00	1,94	1,92	1,86
Humhbb	73.308	2,20	1,97	1,92	1,91	1,92
Humhdabcd	58.864	2,21	2,00	1,94	1,92	1,89
Humhprtb	56.737	2,20	1,97	1,93	1,91	1,90
Humretblas	180.388	2,14	1,95		1,90	1,90
Yeast Cr.III	315.338	2,18	1,96		1,94	1,95
Mpomtcg	186.609	2,20	1,98	1,97	1,96	1,96
Panmtpacga	100.314	2,12	1,88	1,87	1,86	1,86
Chntxx	155.844	2.19	1,96	1,93	1,93	1,93
vaccg	191.737	2,14	1,92	1,90	1,90	1,90
média	129.491	2,19	1,96	1,92	1,91	1,91

Cadeia	Tamanho	Biocompress-2	GenCompress-2	Compressor CDNA	GTAC	CDNA
Humdystrop	38.770	1,93	1,92	1,93	1,81	1,91
Humghcsa	66.495	1,31	1,10	0,95	1,10	0,54
Humhbb	73.308	1,88	1,82	1,77	1,73	1,68
Humhdabcd	58.864	1,88	1,82	1,67	1,70	1,63
Humhprtb	56.737	1,91	1,85	1,72	1,72	1,69
Humretblas	180.388			1,75		1,66
Yeast Cr.III	315.338	1,92		1,94	1,82	1,90
Mpomtcg	186.609	1,94	1,90	1,87	1,78	1,83
Panmtpacga	100.314	1,88	1,86	1,85	1,74	1,81
Chntxx	155.844	1,62	1,61	1,65	1,53	1,30
vaccg	191.737	1,76	1,76	1,81	1,67	1,69
Hehcmvcg	229.354	1,85	1,85		1,74	
média	137.817	1,81	1,75	1,72	1,67	1,60

foram feitas no gene Humretblas, para o qual não há dados do GTAC. Fica pouco confiável uma média calculada sobre amostras não representativas na qual cada teste é feito sobre amostras distintas. Mas a média mostra um valor aproximado que serve de ordem de grandeza (e para muitos algoritmos a ordem de grandeza é a mesma), como uma referência ao algoritmo.

A segunda tabela mostra o resultado dos algoritmos acima descritos. Não inclui o Cfact que ficou em uma terceira tabela junto com o GenCompress. Os resultados dos algoritmos ficaram próximos mas nota-se que a utilização da metodologia LZ não ajudou muito; sua taxa de compressão ficou em torno de 0,1. Os modelos de contextos finitos obtiveram a melhor taxa de compressão em torno de 0,15, o que não é muito em comparação com a obtida para a linguagem natural.

A terceira tabela utiliza seqüências de DNA distintas e em média bem menores que as utilizadas pelos outros algoritmos. Isto dificulta a confrontação dos resultados. Na comparação entre o desempenho do GenCompress-2 e o do Cfact pode-se concluir que o Cfact possui um desempenho similar ao Biocompress-2.

Cadeia	Tamanho	Cfact	GenCompress-1	GenCompress-2
Atatsgs	9.647	1,78	1,67	1,68
Atefla23	6.022	1,58	1,54	1,54
Atrdnaf	10.014	1,81	1,79	1,79
Atrdnai	5.287	1,49	1,42	1,41
Hsg6pdgen	52.173	1,93	1,80	1,81
Xlxfg512	19.338	1,49	1,37	1,38
mmzp3g	10.833	1,91	1,85	1,86
Celk07e12	58.949	1,71	1,62	1,63
média	21.533	1,71	1,63	1,64

Cabe destacar que a média do algoritmo GenCompress-2 variou de 1,75 na segunda tabela, para 1,64 na terceira. Isto se deve exclusivamente a não representatividade das amostras. Porém os dados comparativos da segunda tabela são importantes, pois os algoritmos são testados aproximadamente para os mesmos genes e revelam a grandeza relativa entre eles. As duas primeiras tabelas podem ser consideradas como uma só, já que as cadeias utilizadas foram praticamente as mesmas. Nelas os compressores estão ordenados da esquerda para a direita, em função das taxas de compressão. O CDNA não é um compressor, e portanto apresenta uma estimativa de entropia e não uma taxa de compressão. Foi o melhor valor obtido. O Cfact foi aplicado em seqüências bem menores que os demais algoritmos. Utilizando o GenCompress-2 como base e mantendo a proporção entre as taxas de compressão obtidas por estes algoritmos para os valores da segunda tabela, a taxa de compressão do Cfact ficaria em torno de 1,82. O Cfact ficaria posicionado imediatamente à esquerda do Biocompress-2, nesta tabela. Mas a não representatividade das amostras impede a clara identificação do posicionamento correto do Cfact.

Capítulo 13

Previsibilidade do DNA à Distância

A previsibilidade das cadeias polinucleotídicas sempre tem sido feita para o sucessor imediato. A entropia foi definida desta forma, baseada no fato de que a seqüência segue o processo de Markov (seção 3.2). Porém no DNA nem sempre isto acontece. No processo da síntese protéica, a proteína chamada repressor inibe a produção da nova proteína (seção 2.4). Muitas vezes o gene repressor (que gerou esta proteína) não é vizinho ao gene que ficou inibido na célula. No DNA nem sempre o que está próximo está exercendo influência em certa porção da seqüência.

Na busca de regularidades nas possíveis distâncias de influência, foi criado um estimador de entropias que abrange as distâncias de zero a mil. Foi utilizada a ordem 3 neste estimador. Além de ser o tamanho dos códons permite o manuseio dos 64 antecessores. A distância zero é a normalmente utilizada pelos compressores e estimadores de entropia. Dados os três antecessores imediatos, sabe-se a previsibilidade do próximo símbolo. A distância um indica que nestas condições é definida a probabilidade do símbolo que possui entre ele e seus antecessores exatamente um símbolo (distância). Na distância dois são dois símbolos da cadeia entre o próximo símbolo e seus antecessores. E assim sucessivamente até a distância mil. A distância mil foi escolhida por praticidade computacional. A maioria dos genes é maior que este valor e este tamanho não é suficiente para capturar a motivação deste teste. Entretanto o custo computacional é grande para valores maiores. Este estimador trabalha em tempo $O(N.D)$, sendo N o tamanho da seqüência e D a distância máxima adotada. Desta forma, quando é adotada a distância 1000, para cada nucleotídeo da seqüência lida são feitas 1000 operações com os códons antecessores.

Houve uma opção pelo trabalho com os antecessores contíguos na cadeia e pela utilização da ordem 3. Embora seja possível tratar os antecessores na distância[1], esta opção não foi utilizada, ficando os três antecessores juntos e o sucessor à distância.

[1]Na distância zero, os três antecessores imediatos; na distância um, entre o primeiro e o segundo antecessor exatamente um símbolo; da mesma forma, do segundo para o terceiro antecessor e do terceiro para o próximo símbolo (ordem 3).

A proposta de trabalho deste estimador foi a atenção dentro do gene, na busca da previsibilidade para as distâncias de zero a 1000. Embora seja possível a utilização de algum genoma completo, ou em cromossomos, a tentativa está na pesquisa de alguma distância próxima que influencie a cadeia como um todo.

13.1 Bases de teste

A base de testes utilizada foi pequena e dificilmente será representativa para as cadeias do DNA. Foi feita uma aplicação no uso de diversos tipos de cadeias de DNA, como genomas completos, cromossomos e principalmente genes. Porém não foi cogitado a utilização deste estimador no trato de cadeias com somente íntrons ou somente exons. Todas as cadeias genômicas foram obtidas da Internet e foram obtidas no segundo semestre do ano 2000.

O cromossomo Y do ser humano foi obtido a partir do endereço eletrônico da NCBI [53]. Está dividido em diversas cadeias, sendo escolhida a cadeia NT_001402 que começa na posição 15450 KB deste cromossomo. Esta cadeia, por sua vez, possui oito componentes. O estimador foi executado para cada um deles.

O genoma completo do *Methanococcus jannaschii* também foi obtido a partir do endereço eletrônico da NCBI. Ele foi estudado por Nevill-Manning e Witten [39].

O gene humano da susceptibilidade ao retinoblastoma, o Humretblas, foi a base para o desenvolvimento do estimador de entropias, recebendo medições e testes específicos que não foram aplicados a nenhum dos outros genomas. A escolha deste gene se deve principalmente à quantidade de nucleotídeos de sua cadeia por sua utilização por Loewenstern e Yianilos [31]. Ele é grande o suficiente para ser representativo e pequeno o suficiente para ser testado em alguns experimentos mais demorados.

Foram copiados alguns genes humanos, de roedores, de plantas e de vírus de outro endereço eletrônico [44]. Para que a aplicação do estimador de entropia seja feita sobre um conjunto amplo de genes. Foram destacados o genoma completo do vírus da varíola, o Vvcgaa, com 185.575 nucleotídeos e o Scchriii que é o seqüenciamento completo do cromossomo III do *Saccharomyces cerevisiae*, analisado por Nevill-Manning e Witten [39].

13.2 Resultados

A utilização dos contadores da ordem zero para o gene Humretblas fornece uma estimativa da entropia em torno de 1,95021. Na busca de uma estimativa de entropia à distância para o gene Humretblas, o estimador de entropia foi aplicado para diversas distâncias máximas de

5 a 1.000. Porém a melhor estimativa para a ordem três ficou sempre com a distância zero (previsibilidade usual) com estimativa de 1,89974 bits por símbolo.

Na tabela abaixo, $D_{\max}$ representa a distância máxima para a qual foi computada a estimativa de entropia à distância. A *posição* indica em qual distância ocorreu a estimativa máxima da entropia; o *maior valor* mostra esta estimativa e a *razão* é o resultado da divisão da entropia da distância zero pela da distância máxima.

Estimativas de entropia para o Humretblas em diversas distâncias

$D_{\max}$	Posição	Maior valor	Razão
5	4	1,94034	0,97908
10	10	1,94186	0,97831
25	24	1,94514	0,97666
50	44	1,94584	0,97631
100	96	1,94679	0,97583
200	186	1,94765	0,97540
400	361	1,94863	0,97491
600	520	1,94877	0,97484
800	781	1,94880	0,97483
1000	892	1,94903	0,97471

Aparentemente, a entropia da distância máxima está convergindo para o valor da entropia da ordem zero.

Uma segunda medida feita especificamente para o gene Humretblas foi a entropia para as ordens de um a três. Na ordem zero não há como medir a entropia à distância. Neste cálculo das ordens, as distâncias variaram de 0 a 1.000 e a distância 0 novamente foi a que apresentou a menor estimativa de entropia. Na tabela abaixo, *ordem* é a ordem para a qual foram feitas as estimativas de entropia, $P_{\min}$ representa a distância que apresentou a menor entropia e *menor valor* mostra o menor valor calculado da entropia, $P_{\max}$ a distância que apresentou a maior entropia e *maior valor* este valor. A *razão* é a mesma da tabela anterior.

Estimativas da entropia do Humretblas para as ordens um, dois e três

Ordem	$P_{\min}$	Menor valor	$P_{\max}$	Maior valor	Razão
1	0	1,91385	866	1,95003	0,98145
2	0	1,90501	845	1,94976	0,97705
3	0	1,89974	892	1,94903	0,97471

O aumento da ordem diminuiu a estimativa da entropia tanto para a distância zero quanto para as demais distâncias. Porém a distância zero continua sendo a mais previsível.

Além dos testes com o Humretblas, foram feitas as medições de entropia para diversas seqüências de DNA. Na tabela a seguir "cadeia" indica o nome da cadeia genética utilizada para fazer as estimativas de entropia e "tamanho" o tamanho da cadeia. Os demais campos possuem o mesmo significado que os campos homônimos da tabela anterior.

Estimativas de entropia para diversas cadeias de DNA utilizando ordem três e distâncias até mil

Cadeia	Tamanho	$P_{\min}$	Menor valor	$P_{\max}$	Maior valor	Razão
Humretblas	130.388	0	1,89974	892	1,94903	0,97471
Hsmhcapg	66.108	0	1,92160	861	1,98379	0,96865
Humghcsa	66.495	0	1,91527	906	1,99799	0,95860
Humhbb	73.324	0	1,90777	716	1,96610	0,97033
Humhdabcd	58.864	0	1,91809	829	1,99479	0,96155
Humhprtb	56.737	0	1,90617	726	1,96876	0,96821
Hummmdbc	68.499	0	1,91139	801	1,99613	0,95755
Humneurof	100.849	0	1,89927	984	1,95451	0,97174
Humtcradcv	97.634	0	1,91706	866	1,98903	0,96382
Humtcradcv	55.136	0	1,89396	330	1,94760	0,97245
Média humanos	77.403	0	1,90903	791	1,97477	0,96676
Methanococcus jannaschii	1.664.957	0	1,83781	994	1,89732	0,96864
NC_001732	58.405	0	1,81614	818	1,85376	0,97971
Chmpxx	121.024	0	1,82955	960	1,86430	0,98136
Chntxx	155.844	0	1,92724	984	1,95603	0,98528
Chosxx	134.525	0	1,93410	990	1,96363	0,98496
Clegcga	143.171	0	1,79661	981	1,82519	0,98434
Epfcpgc	70.028	0	1,91496	894	1,94065	0,98676
Mpomtcg	186.608	0	1,95750	884	1,98267	0,98731
Mtpacg	100.314	0	1,85726	427	1,87751	0,98921
Scchriii	315.338	0	1,94319	949	1,96096	0,99094
Yscmtcg	78.294	0	1,55729	993	1,66937	0,93286
Média plantas	145.016	0	1,85752	896	1,89337	0,98033
Mmbgcxd	55.856	0	1,89736	670	1,96769	0,96426
Mustcra	94.647	0	1,92238	535	1,99138	0,96535
Ratcryg	51.391	0	1,92716	991	1,98729	0,96974
Média roedores	67.298	0	1,91563	732	1,98212	0,96645
Asfv55kb	55.098	0	1,93099	412	1,95762	0,98640
Ebv	172.281	0	1,92640	980	1,96855	0,97858
Helcg	152.260	0	1,87422	992	1,90012	0,98637
Hehcmvcg	229.354	0	1,95339	998	1,98425	0,98445
Hevzvxx	124.884	0	1,96732	983	1,99406	0,98659
Hslulr	108.359	0	1,89022	992	1,91247	0,98837
Hsecomgen	153.223	0	1,96684	829	1,98599	0,99036
Hsgend	112.930	0	1,88642	982	1,92650	0,97920
Ihlcg	134.226	0	1,94203	992	1,98667	0,97753
Vaccg	191.737	0	1,89993	961	1,91842	0,99036
Vvcgaa	185.575	0	1,89613	547	1,91144	0,99199
Média virais	147.266	0	1,92126	879	1,94964	0,98547

Não foi calculada a média para o *Methanococcus jannaschii*, pois a primeira seqüência é o genoma completo deste ser. Foram desprezados 13 nucleotídeos deste genoma, pois o estimador de entropia à distância somente prevê os nucleotídeos "a", "c", "g" e "t", mas o contador de onde foi baixado o genoma aponta os seguintes valores: 573.429 "a", 258.665 "c", 264.573 "g", 568.290 "t" e 13 outros nucleotídeos num total de 1.664.970. Desprezando estes 13 nucleotídeos, o genoma fica com 1.664.957. Ele possui uma entropia estimada em 1,89800 bit por nucleotídeo, para o modelo de contextos finitos de ordem zero, um pouco maior que o atingido pelo estimador de entropia à distância, no pior caso.

O estimador também foi utilizado em o segmento NT_001402 do cromossomo Y humano e os resultados obtidos estão tabulados a seguir:

Estimativas de entropia para segmentos do cromossomo Y humano ordem três e distâncias até mil

Cadeia	Tamanho	$P_{\min}$	Menor valor	$P_{\max}$	Maior valor	Razão
NT_001402 AC_002531	197.900	0	1,89415	937	1,94860	0,97206
NT_001402 AC_002992	154.848	0	1,91881	923	1,99085	0,96381
NT_001402 AC_004474	148.280	0	1,90604	967	1,96326	0,97085
NT_001402 AC_004617	176.552	0	1,91380	952	1,98375	0,96474
NT_001402 AC_004772	114.330	0	1,90544	1.000	1,97897	0,96285
NT_001402 AC_004810	86.491	0	1,90391	886	1,96581	0,96851
NT_001402 AC_005942	26.458	0	1,90331	876	1,97778	0,96235
NT_001402 AC_006565	40.879	0	1,89052	984	1,95119	0,96890
Média C Y NT_001402	118.217	0	1,90450	941	1,97003	0,96676

13.3 Conclusões

O cálculo da estimativa da entropia à distância pode seguir duas vertentes: os antecessores, que servem de base para o cálculo das probabilidades dos sucessores, podem estar juntos (utilizada neste trabalho) ou separados. O cálculo feito com os antecessores separados, pela mesma distância que se deseja prever, equivale a um modelo de estados finito, no qual ocorre uma seqüência circular de estados. Porém esta possibilidade não foi aqui explorada.

Cabe destacar que a entropia estimada à distância, com os antecessores juntos, utilizando o modelo de contextos finitos de ordem 3 nas seqüências do DNA, convergiu para o valor da estimativa da entropia calculada pelo modelo de contextos finitos de ordem zero. Isto ocorreu a partir da distância um (a distância zero equivale ao modelo de contextos finitos usual), indicando uma quebra na previsibilidade da cadeia. A convergência foi para o modelo de ordem zero porque não é possível piorar a estimativa da entropia além deste limite, em outras palavras, houve a perda da previsibilidade, à medida que a distância cresceu.

Os testes feitos com o Humretblas mostraram que à medida que a distância aumenta, a

previsibilidade da cadeia diminui. A ruptura da previsibilidade não é imediata e o motivo deste fato não está claro. Mas é viável supor que para as cadeias de DNA o modelo de sua formação pode conter relações de previsibilidade de longa distância. À medida que as distâncias vão aumentando, estas relações perdem influência gradativamente até que a estimativa da entropia reflita somente a quantidade de cada símbolo na seqüência. Outra conclusão para os testes utilizando o Humretblas foi que o aumento da ordem do modelo diminui a incerteza. Há um forte indicativo de que nas cadeias de DNA ocorrem influências locais fortes que em conjunto com as influências à distância retardam ainda mais a queda da previsibilidade à medida que a ordem cresce. Os testes foram feitos até a ordem três. Talvez seja possível capturar alguma ocorrência das influências à distância (se estas existirem) com a utilização de ordens maiores. Como as cadeias de DNA possuem um tamanho relativamente pequeno, a utilização de ordens maiores pode provocar perda da representatividade do processo, invalidando qualquer resultado obtido.

Analisando o resultado do estimador de entropia à distância aplicado às diversas cadeias de DNA, pode-se constatar que a distância zero sempre forneceu a melhor previsibilidade. Os fatores de influência à distância são menos intensos que os fatores de influência locais. Para a maioria das cadeias (35/43), a pior estimativa da entropia ocorreu para distâncias maiores que 800, indicando que a perda de previsibilidade ocorre de maneira gradual. Não foi feita uma análise específica em cada cadeia que apresentou a pior estimativa de entropia em distâncias menores que 800. Porém em 7 destas 8 cadeias o seu tamanho é menor que 101.000. A menor cadeia analisada, cujo tamanho é de 26.458 teve a pior entropia para a distância 876 e é a menor distância para as cadeias analisadas do cromossomo Y humano. Este fato mostra que não apenas o tamanho da cadeia influencia na distância da pior entropia mas que há outros fatores envolvidos; a causa deste fato fica em aberto.

As médias apresentadas nas tabelas anteriores, em alguns casos foi calculada sobre valores discrepantes, como para os valores de $P_{\max}$ dos genes humanos. Visando eliminar as discrepâncias, será mostrada a média obtida. O maior e o menor valor em cada grandeza serão desprezados e será considerado um mínimo de cinco observações para cada medida. Estas médias estão na tabela a seguir e os valores sublinhados nas tabelas anteriores não entraram neste cálculo.

valores médios obtidos a partir do estimador de entropia a distância ordem três e distâncias até mil

Tipos de cadeia	Tamanho	$P_{\min}$	Menor valor	$P_{\max}$	Maior valor	Razão
Genes humanos	73.564	0	1,90935	825	1,97527	0,96691
Methanococcus jannaschii	1.664.957	0	1,83781	994	1,89732	0,96864
Genes de plantas	131.397	0	1,88613	949	1,91261	0,98560
Genes virais	148.386	0	1,92137	918	1,95021	0,98563
Cromossomo Y NT_001402	120.230	0	1,90444	942	1,97013	0,96661

Os valores do *Methanococcus jannaschii* não seguem esta regra, pois se trata da totalidade

de seu genoma e estes valores são significativos para este organismo. Não foi calculada a média para os roedores devido à pouca quantidade de genes testados.

Para as cadeias de DNA analisadas, houve variações individuais relevantes, exceto na grandeza razão. Para esta grandeza, excetuando-se um ou outro caso anômalo, a ordem de grandeza para genes do mesmo grupo foi a mesma. Para os genes humanos os valores variaram de 0,9575 a 0,97471. Quando o grupo é mais restritivo, por exemplo nas divisões do NT_001402 do cromossomo Y humano, estes valores ficaram mais próximos ainda e variam de 0,96235 a 0,97206. No caso específico dos genes humanos, a média ficou muito próxima, quer os extremos sejam levados em conta ou não. Este valor por si não deve ser considerado como um parâmetro isolado, pois os genes virais e os genes das plantas também obtiveram valores bem próximos, apesar de grupos de genomas distintos. Outra grandeza que se mostrou confiável foi $P_{\max}$. No cromossomo Y humano na NT_001402 a menor cadeia de DNA é aproximadamente sete vezes e meia menor que a maior cadeia, mas a distância $P_{\max}$, na qual foi observada a maior estimativa de entropia, variou menos de 15%. Houve casos em que anômalos envolvendo $P_{\max}$, porém esta grandeza pode ser útil na comparação genômica, quando associada à grandeza razão e ao tamanho das cadeias.

Embora a estimativa da entropia à distância não tenha servido para diminuir a incerteza sobre as cadeias de DNA, ela pode ser útil como medida complementar para a comparação genômica. A análise inicial do estimador de entropia à distância indica uma possibilidade de auxílio na busca de similaridade das cadeias genômicas, sendo uma forma complementar de avaliação.

Capítulo 14

Conclusões

Foi feita uma análise ampla a respeito da estimativa da entropia das cadeias do DNA. Alguns aspectos devem ser levados em conta para esta análise. Este trabalho não contou com o acompanhamento de especialistas em código genético. Toda a análise foi feita a partir da representação primária do DNA e não houve considerações a respeito de possíveis significados biológicos existentes. Com a quebra do vínculo entre o significado biológico e os estimadores de entropia neste texto, tratou-se o DNA como um subconjunto de Σ^*, sendo $\Sigma = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}, \}$. A partir dos resultados analisados, buscou-se algum significado ou aplicação prática.

O objetivo principal desta dissertação foi a análise da entropia das cadeias do DNA, a partir dos artigos científicos [19, 31, 39, 28, 17, 10, 13, 29] que abordam este tema, ou temas relacionados. Os experimentos específicos para este trabalho ficaram restritos às aplicações do estimador de entropia à distância para as fitas de DNA. Cada artigo abordou a utilização dos estimadores de entropia nas seqüências de DNA sob um enfoque próprio, individual. Os enfoques distintos foram agrupados para uma melhor análise. O primeiro aspecto discorre sobre a validade da utilização da estimativa da entropia nas seqüências do DNA, vinculando desta forma a entropia com o DNA (seção 14.1). O segundo enfoque trata dos estimadores de entropia aplicados às seqüências de DNA (capítulos 12 e 13). O último descreve a análise das aplicações de estimadores de entropia para as seqüências do DNA (capítulo 11).

14.1 Sobre a entropia do DNA

A entropia é uma propriedade da distribuição de probabilidades (seção 3.4) e uma medida de previsibilidade. Quando Shannon a definiu, ele supôs que o seu cálculo fosse feito em uma cadeia gerada por um processo de Markov (seção 3.2). Para que haja uma distribuição de probabilidades de uma cadeia é necessário que exista uma modelo. A adoção de algum modelo no cálculo da distribuição de probabilidades normalmente está vinculado à estrutura utilizada

para a codificação da cadeia (modelo gerador). Porém o modelo gerador das cadeias de DNA, se existir, é desconhecido. Na falta deste, são utilizados para as cadeias de DNA os mesmos modelos da teoria da informação (capítulo 4).

O DNA contém informação. A teoria da informação utiliza modelos criados para a linguagem natural. Porém o modelo adequado para o DNA não é conhecido. Como os modelos utilizados nas cadeias de DNA não refletem a realidade do DNA, eles apenas geram uma distribuição de probabilidades, para a qual é calculada a entropia, um número. Este número está mais ligado ao modelo utilizado do que à seqüência na qual foi aplicado, pois ele está desvinculado do modelo gerador. Tanto Farach et al. [19] quanto Loewenstern e Yianilos [31] deixam bem claro que entropia não está definida para o DNA. Sobre a entropia do DNA, título desta seção, é relevante destacar que a entropia não está caracterizada para o DNA. O processo envolvido na geração (como mensagem que contém alguma informação) do DNA é desconhecido. A entropia é uma forma de tentar, a partir da avaliação de uma grandeza (entropia), a identificação das propriedades ou características específicas do DNA, como conteúdo de informação (sintática). Desta forma, transcendendo para a informação semântica e pragmática, é feita a tentativa de inferência sobre o significado das características sintáticas encontradas.

14.2 Estimadores e compressores aplicados ao DNA

A taxa de compressão das cadeias genômicas é muito menor que a obtida pelos mesmos compressores aplicados à linguagem natural. Na tentativa de melhorar a compressão das fitas de DNA, os pesquisadores utilizaram abordagens alternativas, implementando em seus algoritmos características biológicas ou alterando o enfoque tradicionalmente empregado na teoria da informação. A melhora na taxa de compressão foi pequena, mas positiva. As principais abordagens adotadas foram a comparação inexata e a transformação gramatical.

A comparação inexata foi utilizada por Loewenstern e Yianilos [31] através da distância de Hamming. O CDNA por, eles criado, obteve a melhor estimativa de entropia para as cadeias de DNA. A adoção de pesos auxiliam na previsibilidade do próximo símbolo nos casos em que o contexto ainda não ocorreu. Nevill-Manning e Witten [39] fizeram o CP através de uma generalização do *blending*. Não conseguiram comprimir as proteínas, mas obtiveram uma pequena melhora em suas taxas de compressão. Chen, Kwong e Li [10] implementaram a comparação inexata através de operações de edição. O GenCompress obteve a melhor taxa de compressão utilizando a tecnologia LZ.

A transformação gramatical foi aplicada nas seqüências de DNA, por Lanctot, Li e Yang [28], antes do emprego de um compressor aritmético de ordem zero. O resultado obtido foi muito bom, o melhor compressor para as cadeias genômicas. Com este método foi possível capturar

todas as pequenas repetições existentes nas seqüências de DNA.

A abordagem utilizada neste trabalho foi a estimativa de entropia à distância. Foram feitas considerações sobre suas formas de estimativa, porém não houve sucesso na tentativa de diminuição da incerteza. A estimativa da entropia à distância convergiu para o valor da distribuição de probabilidades do modelo de contextos finitos de ordem zero, apesar da utilização do modelo de ordem três.

É possível que uma melhor estimativa de entropia do DNA seja atingida. O ideal é o trabalho em conjunto com algum especialista em DNA, permitindo a utilização de diversos fatores biológicos como as repetições invertidas e a forma da espiral. O desconhecimento do modelo gerador do DNA compromete sua previsibilidade e a correta estimativa de sua entropia.

14.3 Aplicações práticas da entropia do DNA

Algumas aplicações biológicas tiveram auxílio dos estimadores de entropia para seqüências do DNA: a detecção do limite íntron/exon, a comparação genômica e a construção de árvores filogenéticas.

Farach, Noordewier, Savari, Shepp, Wyner e Ziv [19] encontraram padrões que detectam o início dos íntrons. Embora não tenha havido sucesso na detecção do final dos íntrons, houve uma aplicação prática da entropia para fins genéticos. Eles propuseram que os íntrons podem estar de alguma forma relacionados com os exons que os sucedem. Nenhum dos demais artigos analisados confirmou ou refutou esta hipótese. Provavelmente esta questão ainda está em aberto.

Chen, Kwong e Li [10] criaram um compressor que foi aplicado na comparação genômica. Uma das principais dificuldades da construção de árvores filogenéticas está na definição da proximidade das cadeias genômicas. Utilizando o compressor e outros critérios de decisão, eles recriaram duas árvores filogenéticas.

A vida possui uma grande diversidade. Talvez não seja viável a utilização de alguma grandeza matemática (a entropia, por exemplo) isolada para caracterizar os processos biológicos. A entropia se mostrou útil em alguns casos; em outros fracassou. O DNA possui informação e a entropia é uma das grandezas da teoria da informação. Aos biólogos falta o conhecimento matemático. Aos matemáticos falta o domínio da biologia. Mesmo nos casos nos quais a entropia não obteve sucesso, ela pode ser utilizada em conjunto com outras grandezas matemáticas e outros princípios biológicos para ajudar a desvendar o que está ocultado pelo código genético. O trabalho em conjunto entre biólogos e matemáticos talvez seja o melhor caminho para o aprimoramento destas aplicações.

14.4 Conclusões finais

O DNA possui mais repetições próximas do que seria esperado por uma distribuição aleatória. Processos que exploraram esta característica, direta ou indiretamente, conseguiram diminuir a estimativa da entropia das cadeias do DNA, com destaque para a comparação inexata e para as transformações gramaticais. Ambas pareceram promissoras quando aplicadas aos genomas.

Outra característica do DNA explorada foi a seqüência complementar invertida ou repetições invertidas. Alguns artigos utilizaram esta característica [31, 28, 10], mas não obtiveram uma taxa de compressão relevante.

As proteínas são muito pouco previsíveis. Embora ainda não haja uma explicação para este fato, é possível que na evolução das espécies somente tenham sobrevivido aquelas que minimizaram o custo da síntese protéica. A substância mais abundante nas células é a água. Se ela for retirada análise, sobram os demais componentes da célula. As proteínas respondem por mais da metade do peso destes componentes. Para obter esta maior quantidade de matéria e energia, devem ser ingeridos mais alimentos. Pela teoria da seleção natural, os indivíduos (espécies) que utilizarem menos alimento sobrevivem (melhor adaptação ao hábitat). Talvez por este fato, as tentativas de compressão das seqüências de proteínas não são efetivas.

O processo de geração do DNA não segue o processo de Markov. Como prova desta alegação são citadas as seqüências com repetições invertidas. A previsibilidade deste segmento depende da necessidade biológica de ligação destas subcadeias nas fitas de RNA, transcritas a partir desta seqüência (estruturas cruciformes seção 2.6). Assim, a previsibilidade de um destes segmentos depende do outro segmento de mesmo tamanho, que não é contíguo a este. Por não ser contíguo sua previsibilidade não segue o processo de Markov.

O modelo gerador do DNA não é conhecido. As proteínas provavelmente já estão otimizadas pela natureza. O DNA não segue o processo de Markov. São estas as condições de contorno que envolvem a estimativa da entropia do DNA. A entropia não está caracterizada para o DNA e ainda não se mostrou uma ferramenta efetiva no auxílio dos biólogos. Mas a entropia possui um potencial para ser útil na análise das seqüências de DNA. Pode servir, tanto para caracterizar regiões que apresentem certas propriedades, quanto na análise de regiões com significado biológico conhecido. Esta é a razão para a utilização de estimadores de entropia para seqüências de DNA.

Glossário

acesso direto forma de acesso que permite ir diretamente a determinada posição de um arquivo (em disco) ou endereçamento da memória. Normalmente é necessário criar uma estrutura de dados, chamada índice, para viabilizar o acesso direto.

ácido desoxirribonucléico vide DNA.

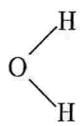
ácido graxo é o lipídeo mais simples. Serve para constituir os lipídeos mais complexos.

ácidos nucléicos são macromoléculas ácidas. São polímeros lineares de nucleotídeos. Há dois tipos de ácidos nucléicos: ácido desoxirribonucléico e ácido ribonucléico.

ácido ribonucléico vide RNA.

adaptação ajustamento de um organismo às condições do meio ambiente.

água substância química composta por dois átomos de hidrogênio e um átomo de oxigênio (H_2O). É a substância encontrada em maior quantidade nos seres vivos [22]. A vida é impossível sem água no estado líquido [50]. A maior parte das reações químicas que se passam no interior dos organismos exige água [50]. Os processos fisiológicos se produzem exclusivamente em meio aquoso [58].

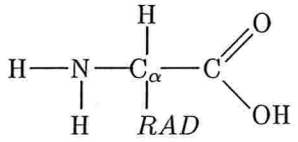


aleatoriedade imprevisibilidade.

alfabeto conjunto de símbolos utilizados em uma mensagem. Normalmente designado pela letra grega Sigma (Σ).

algoritmo seqüência finita de ações que devem ser executadas para o cumprimento de uma tarefa. Normalmente é ligada à matemática (cálculo da raiz quadrada). Em computação, algoritmo é a seqüência finita de operações do computador necessária para a execução de um processamento de dados.

aminoácidos moléculas que formam as proteínas. Existem 20 aminoácidos. Cada um é representado por três nucleotídeos no código genético. A maioria dos aminoácidos possui mais de um códon que o representa. Possuem a estrutura química abaixo, onde *RAD* representa o radical que distingue um aminoácido do outro, e C_α o carbono principal do aminoácido. As fórmulas químicas da fenilalanina e glicina estão descritas mais à frente.



arquebactérias bactérias primitivas.

árvores filogenéticas são as representações através das árvores da história evolucionária das espécies.

auto replicação processo que caracteriza os seres vivos. Capacidade de reprodução dos seres vivos.

autotrófico vide seres autotróficos.

bactéria organismo procariótico unicelular.

base nitrogenada composto cíclico contendo nitrogênio. Neste texto, seu significado se confunde com o de nucleotídeo, embora seja somente um componente deste.

base pirimidínica base nitrogenada derivada da pirimidina, também chamada de base pirimídica. Citosina, timina e uracil

base purínica base nitrogenada derivada da purina, também chamada de base púrica. Adenina e guanina.

biologia molecular "Nós também valorizamos o fato de que a biologia molecular não trata dos aspectos triviais dos sistemas biológicos. Ela está no cerne da questão. Quase todos os aspectos da vida são construídos no âmbito das moléculas e sem compreendê-las nós podemos ter apenas uma visão muito superficial da própria vida. Todas as abordagens em níveis superiores são suspeitas até serem confirmadas no âmbito molecular." Francis Crick, *"What Mad Pursuit"*, 1988.

biosfera conjunto de todos os ambientes naturais da Terra nos quais os seres vivos vivem e se reproduzem [50].

bit menor unidade da memória do computador. Recebe sinais elétricos e pode estar ligado ou desligado. O bit ligado é representado por 1 e o desligado por 0. Os computadores digitais utilizam códigos binários pelo fato do bit possuir apenas estes dois estados. Componente do byte.

blending palavra da língua inglesa que significa mistura. No contexto de modelos e técnicas de compressão, *blending* é a utilização de diversas ordens do modelo de contextos finitos.

byte menor unidade endereçável da memória do computador. Normalmente um byte possui o tamanho de oito bits. É através do byte que são implementados os códigos (ASCII, EBCDIC, etc) para representação dos símbolos do quotidiano. Alguns códigos, como o UNICODE, necessitam de dezesseis bits para serem implementados, utilizando dois bytes em sua representação.

cadeia alimentar é a passagem de energia e matéria do meio ambiente para os seres vivos, advindas do meio aos produtores e passando destes para os consumidores. O retorno da matéria dos seres vivos para o meio ambiente é feito através dos decompositores.

cadeia genômica ou cadeia de DNA, vide fita.

cadeia polinucleotídica seqüência de nucleotídeos cuja união é sempre feita entre o fosfato de um nucleotídeo e a pentose do nucleotídeo vizinho.

caractere símbolo utilizado na linguagem natural (em especial no alfabeto romano), composto por letras (com ou sem acentuação), números, sinais e abreviações (etc representa a expressão do latim *et cetera*, etc) e outros símbolos ($\int$ nas expressões matemáticas, @ em expressões comerciais, etc).

carboidratos são compostos orgânicos da forma $(CH_2O)_n$, onde n é o número de átomos de carbono. É um tipo de açúcar.

celacanto "peixe ósseo pulmonado, considerado um dos predecessores dos anfíbios. Conhecido em registros fósseis, o celacanto era tomado por extinto até 1938, quando uma espécie foi descoberta vivendo em águas profundas da costa sul-africana. A confirmação de muitas características propostas por pesquisadores a par-

tir de fósseis, quando da descoberta do animal vivo, causou grande impacto nos meios científicos. Os espécimes têm cerca de 1,5m de comprimento e escamas circulares sobrepostas" [41]. É o mais famoso fóssil vivo [51].

célula compartimento vivo composto por uma membrana celular, o citoplasma e contendo o DNA do ser vivo do qual ela é parte (ou o próprio ser vivo se este for unicelular).

célula especializada célula de seres eucarióticos que possui uma função específica. Por exemplo: neurônio, célula cardíaca, célula reprodutora (ou gameta), etc.

célula somática célula que possui dois jogos de cromossomos.

citoplasma região interna das células, onde ficam as organelas, incluindo o núcleo celular. É onde ocorrem a maioria das reações químicas necessárias às células.

classes de condição é o conjunto de probabilidades condicionais dada a ocorrência de algum evento. Para os modelos de contextos finitos, para cada contexto há uma classe de condições.

cloroplasto organela que realiza a fotossíntese.

codificação é o processo de representação da informação através de uma mensagem ou a modificação da forma de representação da mensagem, sem alteração do conteúdo.

codificador aquele que faz a codificação ou compressão. Compressor.

codificadora ou cadeia ou seqüência codificadora é a cadeia polinucleotídica utilizada na representação primária do DNA. Vide fita.

código conjunto de regras utilizadas para a representação de alguma informação em uma mensagem.

código binário forma de representação de uma informação ou mensagens sobre o alfabeto $\{0, 1\}$. Por exemplo: EBCDIC, ASCII, etc.

código genético é a representação das proteínas através dos polinucleotídeos. O DNA (em alguns casos o RNA) de uma espécie.

códon conjunto de três nucleotídeos que representam um aminoácido (ou algum dos símbolos de final de codificação) no código genético.

comensalismo é a relação na qual alguns dos seres envolvidos dela tiram proveito sem causar lucros ou prejuízos aos demais participantes. Por exemplo um peixe que acompanha tubarões e come as sobras de suas caçadas.

comparação genômica ação que visa estimar as proximidades entre as cadeias de DNA, por exemplo, para fins da feitura de árvores filogenéticas.

complexidade dos algoritmos informalmente: relação entre algumas das entradas do algoritmo e o número de operações executadas para seu processamento. Pode ser analisada individualmente, na média, ou para o pior caso (limite superior).

complexo de Golgi organela envolvida na modificação e no transporte de moléculas formadas no retículo endoplasmático.

compressão codificação com o objetivo de tornar a mensagem menor.

compressão ambígua feita por algoritmo de compressão que possibilita mais de uma decodificação da mensagem comprimida, dificultando a decodificação da mensagem original.

compressão aritmética técnica de compressão que transforma a mensagem em um número no intervalo $[0, 1)$.

compressão de dados algoritmo de compressão utilizado para comprimir dados diversos: textos, figuras, etc.

compressão irreversível técnica de compressão que não permite o retorno à mensagem original. Não utilizada na compressão de textos, nem nas seqüências genômicas.

compressão reversível técnica de compressão que permite o retorno à mensagem original.

compressibilidade propriedade de alguma mensagem de ser compressível. Taxa de compressão: o quanto é compressível .

compressor algoritmo utilizado para fazer a compressão.

compressor LZ compressor que segue os princípios criados por Ziv e Lempel.

compressor universal algoritmo que pode ser utilizado em qualquer tipo de dado. O código Morse não é universal, pois não pode transmitir músicas ou imagens.

comunicação processo de representar a mensagem em um local distinto.

comunidade é o agrupamento de indivíduos de espécies diferentes, convivendo em um mesmo local.

consumidores primários são seres vivos que se alimentam dos produtores.

consumidores quaternários são os seres que se alimentam dos consumidores terciários.

consumidores secundários são os seres que se alimentam dos consumidores primários.

consumidores terciários são os seres que se alimentam dos consumidores secundários.

criacionismo teoria pela qual o mundo e a humanidade foi criado por Deus, conforme relatado no livro bíblico do Gênesis [41]. Esta teoria nega as teorias da geologia, da biologia e da física sobre a criação do mundo, vida e universo. Teoria da origem dos seres por criação, oposta à evolução expontânea [14].

cromatina é um componente nuclear (da célula) que, ao microscópio óptico, aparece na forma de uma mancha irregular, quando tratada com um corante básico como o azul de metileno.

cromossomo constituído por um filamento de cromatina dobrado várias vezes sobre si mesmo, adquirindo assim um aspecto compacto, em forma de bastonete. Cada filamento de cromatina é formado por uma longa molécula de DNA que, em algumas regiões, aparece enrolada em volta de proteínas chamadas *histonas.*

cromossomos homólogos os cromossomos existem aos pares nas células somáticas. Os cromossomos que formam cada par são chamados homólogos. Nos seres humanos há 22 pares de cromossomos homólogos e um par de cromossomos sexuais. Somente um cromossomo de cada par é passado para os gametas.

custos da compressão bits utilizados pelo processo de compressão, além dos bits necessários para a representação da mensagem.

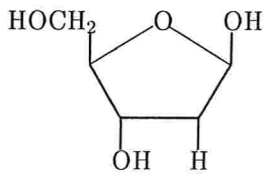
decodificação é o processo inverso da codificação ou compressão.

decodificador aquele que reverte a compressão ou codificação.

decompositores são os seres que transformam a matéria orgânica em matéria inorgânica.

desnaturação processo pelo qual são separadas as duas cadeias polinucleotídicas do DNA, apenas em uma região específica; tem por finalidade a exposição de uma delas para que seja feita a sintetização do RNA mensageiro.

desoxirribose pentose do ácido desoxirribonucléico.



desvio padrão é a raiz quadrada da variância.

dicionário técnica de compressão que consiste na criação de frases que serão usadas como entradas no dicionário. A codificação é feita pela representação do código da entrada no dicionário em substituição à frase.

dígito qualquer um dos algarismos arábicos do 0 ao 9.

digrama conjunto de duas letras ou palavras, conforme o contexto utilizado.

distância de Hamming é o número de posições para as quais duas cadeias de mesmo tamanho possuem conteúdos distintos.

distribuição de probabilidades uma distribuição na qual a soma das probabilidades de todos os eventos é 1. A distribuição será **discreta** se o número de eventos for finito; será **uniforme** se os eventos forem equiprováveis.

divisão celular processo de criação de novas células (em seres multicelulares), reprodução (em seres unicelulares) e eventualmente a produção de gametas. No processo de divisão celular, o DNA sai do núcleo das células e é duplicado.

DNA ácido desoxirribonucléico. Dupla cadeia polinucleotídica, que contém toda a informação necessária aos processos da vida.

ecologia é o estudo das relações dos seres vivos entre si e com o meio em que vivem.

ecossistema sistema de relações entre o meio ambiente e os seres que nele vivem. Existe troca de energia e matéria entre eles. Os fatores bióticos e abióticos são relevantes num ecossistema.

entrada no dicionário cada frase escolhida para ser representada por um código (ou índice).

entropia é a medida da quantidade de informação de Shannon. Está vinculada à uma distribuição de probabilidades fornecida por um modelo. Grandeza que representa o limite de compressibilidade. Normalmente representada em bits ou bits por símbolo.

entropia condicional é a redução da incerteza oriunda da ocorrência de algum evento. Nestas circunstâncias, a distribuição de probabilidade, para a qual é calculada a entropia é condicionada à ocorrência deste evento.

enzima proteína com propriedades de catalisar (acelerar) uma reação química sem se consumir.

espaço amostral é o conjunto de todos os resultados possíveis de um evento aleatório.

espécie categoria sistemática que inclui os indivíduos com as mesmas características funcionais e morfológicas e com a capacidade de se reproduzirem entre si. Na nomenclatura binária de Lineu é a segunda denominação. [18] No exemplo *Homo erectus* o *erectus* indica a espécie.

esperança matemática é o valor médio de uma variável aleatória, se esta for finita ou convergir absolutamente.

estimadores de entropia algoritmos que, a partir de algum modelo, calculam a entropia para alguma mensagem.

estimativa da entropia valor calculado, pelo estimador de entropia, para uma mensagem, a partir de um modelo.

estrutura primária do DNA é a representação do DNA pelas bases nitrogenadas na ordem em que estas estão definidas na codificadora. Vide fita.

estrutura primária da proteína é a seqüência de aminoácidos que formam a cadeia peptídica.

estrutura quaternária da proteína refere-se à disposição das subunidades protéicas que formam a molécula.

estrutura secundária do DNA trata dos tipos de helicoidais formadas pelo DNA na célula.

estrutura secundária da proteína trata da organização espacial dos aminoácidos próximos na cadeia peptídica.

estrutura terciária do DNA está ligada ao superenrolamento do DNA.

estrutura terciária da proteína refere-se ao enovelamento da cadeia peptídica.

eucariótico tipo de ser vivo cuja(s) célula(s) apresenta(m) um núcleo protegido pela membrana nuclear.

evento qualquer subconjunto do espaço amostral.

eventos independentes são dois eventos A e B com a seguinte propriedade: $P(A \cap B) = P(A) \times P(B)$. **Eventos mutuamente independentes** são um grupo de eventos independentes, se tomados dois a dois.

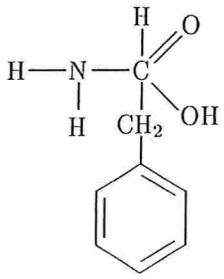
exons regiões dos genes que contêm a codificação dos aminoácidos que serão utilizados na síntese protéica e se intercalam com os íntrons. É a porção do transcrito primário (veja transcrição) que está presente no DNA, mantida no RNA após seu processamento (do RNA).

experimentos aleatórios ou não determinísticos são aqueles que podem ser repetidos indefinidamente sobre as mesmas condições.

fatores abióticos (do ecossistema) fatores desprovidos de vida mas que a influenciam, como a luminosidade, umidade, pressão, temperatura etc.

fatores bióticos (do ecossistema) os seres vivos que habitam o ecossistema.

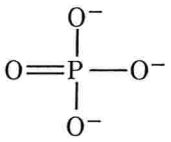
fenilalanina (Phe, F), aminoácido cuja fórmula química é:



fenótipo é a expressão observável de um genótipo.

fita cadeia de DNA (ou genômica), codificadora, seqüência genômica (ou de DNA) e estrutura primária do DNA possuem o mesmo significado: a representação do DNA.

fosfato é o radical livre do ácido fosfórico



fotossíntese processo típico dos organismos vegetais que utiliza a energia da luz para a retirada do carbono da atmosfera (gás carbônico); gera glicose e libera oxigênio de volta para o ar.

frase cada entrada no dicionário.

fraseamento processo que escolhe, a partir da mensagem, quais serão as estradas no dicionário.

freqüência relativa é a razão entre o número de ocorrências de um evento e o número de repetições do experimento.

gametas células especiais para a reprodução sexuada. São tipicamente haplóides, com um único conjunto de cromossomos. A união de dois gametas forma a célula ovo ou zigoto.

gene é a unidade básica da hereditariedade. Região contígua do DNA que contém as informações necessárias para a produção de uma proteína. Nos seres eucarióticos, os genes podem estar ativados ou desativados, de acordo tanto com a função que a célula exerce quanto com influências externas.

genes alelos são genes que ocupam a mesma posição (locus) em cromossomos homólogos.

genoma constituição genética total de um indivíduo.

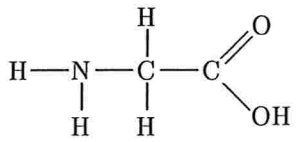
genótipo é a constituição genética de um organismo.

geocentrismo teoria na qual a Terra é colocada como o centro do universo. Ptolomeu (Claudius Ptolemaeus), astrônomo grego do século II da era cristã, defendeu-a em suas obras e ela teve grande autoridade na Idade Média. No século XVI, Nicolau Copérnico, astrônomo polaco, publicou o livro *De revolutionibus orbium coelestium* (sobre as revoluções das esferas celestes) no qual refutou as idéias de Ptolomeu e propôs a teoria do *heliocentrismo*, que coloca o sol como o centro do universo.

geração expontânea teoria pela qual alguns tipos de vida podiam surgir totalmente formados a partir da matéria inanimada [7].

glaciação Congelamento periódico de certas regiões da Terra [46].

glicina (Gly, G), aminoácido cuja fórmula química é:



hábitat é o ambiente natural em que vive uma espécie.

hereditariedade fenômeno de continuidade biológica pelo qual as formas vivas repetem suas características nas sucessivas gerações.

incerteza medida de aleatoriedade. Quanto menor a incerteza de alguma mensagem, maior sua previsibilidade.

informação o mesmo que informação sintática.

informação hereditária ou genética, características transmitidas principalmente através dos processos de reprodução e síntese protéica. Contém a informação necessária para a manutenção e procriação dos seres vivos. Normalmente fica contida no DNA.

informação pragmática forma de tratar o conteúdo da mensagem em relação aos efeitos práticos por ela causados.

informação semântica forma de tratar o conteúdo da mensagem em relação a seu significado, sem que sejam levados em conta seus efeitos (pragmática) ou sua forma (sintática).

informação sintática modo de tratar a mensagem pelas características de sua codificação, forma que a mensagem está representada.

interferência erro no processo de transmissão da informação, que pode ocorrer em qualquer parte deste processo.

íntrons regiões dos genes que se intercalam com os exons, cujas funções ainda não são conhecidas. Parte do transcrito primário (veja transcrição) que não é incluída na forma madura do RNA, após o processamento deste.

íon átomo ou molécula carregada eletricamente, como Cl^-, Na^+ e K^+. Normalmente os íons são solúveis em água.

janela memória utilizada nos algoritmos LZ. É composta por duas partes, a janela a codificar e a janela anterior (já codificada).

junk DNA vide lixo genético.

ligação peptídica nome da ligação dos aminoácidos em uma proteína.

limite íntron/exon problema da detecção dos limites entre os íntrons e os exons nos genes dos seres eucarióticos.

linguagem natural forma de expressão utilizada pelo ser humano, principalmente a escrita e a fala.

lisossomo organela ligada à digestão celular.

livro código representação, através dos modelos semi-adaptativos, das ocorrências mais freqüentes de alguma mensagem. Pode ser uma distribuição de probabilidades (modelos probabilísticos) ou um fraseamento (técnicas de dicionários).

lixo genético região do DNA que não representa nenhum gene. Nos seres eucarióticos a região intergenes (junk DNA) normalmente é bem maior que os genes.

média é a razão da soma dos valores pela número total de observações [33].

mediana é o valor que ocupa a posição central dos dados ordenados [33].

meio de transmissão é o local onde é feita a transmissão. Nos processos da vida, o meio de transmissão normalmente é a célula.

membrana celular estrutura que separa citoplasma do ambiente externo à célula.

membrana interna estrutura que separa as organelas do citoplasma da célula.

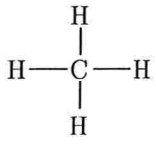
membrana nuclear estrutura que separa o núcleo do citoplasma das células dos eucarióticos.

memória capacidade de guardar eventos ocorridos anteriormente. Nos processos de compressão: quantidade de bits reservada para o que já foi codificado e que pode ser utilizada pelo processo de compressão em andamento.

mensagem alguma forma de representação da informação.

metabolismo conjunto de todas as transformações químicas do organismo [22].

metano (CH_4), substância química gasosa e incolor.



migração entrada ou saída de indivíduos (ou espécies) de um hábitat.

minerais compostos inorgânicos utilizados pelas células. Normalmente aparecem como íons, dissolvidos na água, formando cristais ou combinados com as moléculas orgânicas.

minivulcão estrutura geológica que ocorre normalmente a profundidades maiores que 2.000m onde as placas tectônicas se afastam. É uma forma geológica muito instável que fornece calor e minerais para o ecossistema dos ventos hidrotermais.

mitocôndria organela especializada no metabolismo oxidativo (respiração da célula).

modelo maneira pela qual a distribuição de probabilidades de uma mensagem é obtida.

modelo adaptativo modelo caracterizado pela alteração da distribuição de probabilidades à medida que a mensagem é codificada.

modelo adotado é o utilizado para fornecer alguma distribuição de probabilidades a partir de uma mensagem.

modelo de contextos finitos fornece a distribuição de probabilidades a partir de um certo número fixo de antecessores do próximo símbolo, a **ordem do modelo**.

modelo de estados finitos fornece a distribuição de probabilidades para cada estado existente na mensagem. Desta forma, a probabilidade do próximo símbolo depende tanto de seus antecessores quanto do estado que estes estão.

modelo ergódico é aquele no qual, a partir de cada estado, qualquer outro estado pode ser atingido, se a mensagem for suficientemente longa.

modelo estático possui pré-definidas as distribuições de probabilidades e é independente da mensagem para a qual ele for aplicado. O código Morse é um exemplo clássico de modelo estático.

modelo gerador é um algoritmo que possui alguma distribuição de probabilidades e a partir desta gera uma mensagem. Se a mensagem for suficientemente longa, ela refletirá a distribuição de probabilidades do modelo gerador, se este for ergódico. O processo de Markov é um exemplo de modelo gerador.

modelo gramático possui alguma gramática a ele associada. A probabilidade do próximo símbolo depende dos antecessores e da regra gramatical a qual ele pertença.

modelo inicial existente tanto no codificador quanto no decodificador. Atua no início dos processos adaptativos fornecendo a distribuição de probabilidades. É o utilizado pelo compressor adaptativo antes da leitura do primeiro símbolo da mensagem.

modelo semi-adaptativo lê toda a mensagem e cria o livro código com a distribuição de probabilidades (que será utilizado para codificar e decodificar a mensagem).

molécula é formada pela ligação dos átomos entre si. A força que mantém os átomos unidos é chamada de ligação química. A molécula do gás carbônico (CO_2) é formada por dois átomos de oxigênio e um de carbono.

$$O{=}C{=}O$$

monômeros são pequenas moléculas que servirão de base para a formação dos polímeros biológicos. O mesmo que resíduos.

monossacarídeos são os açucares mais simples, que não podem ser quebrados em glicídios menores. Eles são classificados conforme o número de átomos de carbono que possuem: trioses, tetroses, pentoses e hexoses. A ribose e a desoxirribose são exemplos de pentoses.

mutação genética modificação na seqüência de nucleotídeos do DNA. É **viável** quando permite que o indivíduo continue sobrevivendo, apesar da mutação genética, e **inviável**, caso leve à morte. É **positiva** quando melhora a adaptação do indivíduo ao meio e **negativa** quando piora. As mutações genéticas ocorrem ao acaso (são aleatórias) e podem ocorrer tanto em células somáticas quanto em células germinativas (ou reprodutoras).

mutualismo é a relação na qual seres que teriam dificuldade de viver isoladamente se juntam e promovem uma colaboração mútua. Um exemplo é a relação entre o feijão, produtor de açúcar, e um tipo de bactéria, assimiladora de nitratos. Nesta, o feijão alimenta as bactérias e estas fornecem os sais minerais que ele necessita para seu desenvolvimento.

núcleo existente somente nos seres eucarióticos, possui algumas funções entre as quais se destaca conter em si o DNA, exceto durante o processo de divisão celular.

nucleóide existente somente nos organismos procarióticos, tem funções equivalentes ao núcleo celular (dos eucarióticos), mas deixa o DNA exposto ao citoplasma pela ausência da membrana nuclear.

nucleotídeos são monômeros compostos por três substâncias químicas: base nitrogenada, fosfato e pentose.

ordem do modelo a quantidade de símbolos que são utilizados nos modelos para a previsibilidade do próximo símbolo.

organelas são regiões definidas dentro das células; são separadas do citoplasma por membranas internas e realizam funções específicas.

organismos autotróficos vide seres autotróficos.

oscilação genética (deriva genética) característica de pequenas populações, se dá quando modificações ao acaso produzem alterações na freqüência genótipa. Um exemplo é o princípio do fundador. Quando uma pequena população começa a habitar um meio ambiente. Esta possuirá uma pequena variação genética até que as mutações provoquem um aumento desta variação. O princípio do fundador parece ser o método mais comum de dispersão de inúmeras espécies de plantas e animais.

ovo o mesmo que zigoto.

palavra conjunto de símbolos consecutivos que não apresenta sinais de pontuação, espaço e suas variações.

panspermia teoria pela qual a vida existe em diversos lugares e se reproduz onde encontra um ambiente favorável.

parasitismo é uma relação na qual um organismo, o parasita, se hospeda no corpo de um ser vivo de outra espécie, o hospedeiro, e dele tira os nutrientes necessários para sua vida, com prejuízos ao hospedeiro que pode ficar doente ou até mesmo morrer. Um carrapato, por exemplo.

partição do espaço amostral um conjunto de eventos é uma partição de algum espaço amostral nas seguintes condições: a união de todos os eventos formar exatamente o espaço amostral, não existe o evento vazio e a interseção de quaisquer dois eventos é vazia.

pentose monossacarídeo com cinco átomos de açúcar.

pequenas moléculas possuem em torno de 50 átomos. Com características estruturais peculiares que servem para a formação dos polímeros biológicos ou como substratos e produtos de vias metabólicas.

pirâmide alimentar é a representação da quantidade de seres de cada espécie da cadeia alimentar. Por exemplo a quantidade de plantas, gafanhotos, sapos, cobras e gaviões que formam uma cadeia alimentar.

placas tectônicas são placas semi-rígidas que derivam umas em relação às outras e formam a litosfera (camada superior do manto e a crosta terrestre).

polímeros biológicos são compostos pelos monômeros e são de três tipos: polissacarídeos, proteínas e ácidos nucléicos.

polinucleotídeo (cadeia polinucleotídica) longo filamento composto por nucleotídeos. Representa tanto as cadeias DNA quanto as de RNA.

população é um agrupamento de indivíduos de uma mesma espécie convivendo em uma área comum.

ppm sigla em inglês para previsão por comparação parcial; é utilizada pela técnica de *blending* para decidir o processo do escape.

pré-biótico anterior à vida.

predatismo é a relação na qual um ser vivo, o predador, ataca, mata e come uma caça (presa) inferior a ele na cadeia alimentar. Por exemplo o tamanduá e as formigas.

pressão atmosférica é a pressão exercida pela massa de ar que constitui a atmosfera. No século XVII, Evangelista Torricelli descobriu sua existência.

previsibilidade é a qualidade do que se pode prever. Quanto maior a previsibilidade, menor a quantidade de informação (entropia) da mensagem.

príon agente infeccioso causador da "doença da vaca louca". É uma forma alterada de uma proteína presente no cérebro dos vertebrados. O príon se combina com a proteína normal, alterando sua forma e provocando seus efeitos [22].

probabilidade é o limite da freqüência relativa, quando o experimento é repetido infinitas vezes.

probabilidade condicional é a probabilidade de ocorrer algum evento, dado que um outro evento ocorreu.

probabilidade conjunta é a probabilidade de dois eventos ocorrerem simultaneamente.

procariótico tipo de ser vivo cuja célula apresenta um nucleóide sem a presença da membrana celular.

produtores são os seres autotróficos.

programa forma de codificação de algum algoritmo para que ele seja executado em algum computador.

proteína macromolécula responsável pelo trabalho celular. Formada por uma cadeia de aminoácidos.

protozoários animais eucarióticos unicelulares.

quantidade de informação independe do tamanho da mensagem. Incerteza. Medida através da entropia.

recepção operação de recebimento da mensagem, normalmente seguida da decodificação (ou nova codificação) para a utilização desta (aspecto pragmático). Na síntese protéica, a recepção é feita pelos ribossomos, que efetivam a tradução do mRNA para uma proteína.

recepção incremental processo em que cada bit é transmitido quando é definido. É utilizada por alguns codificadores.

receptor destino da mensagem. Em muitos casos é também o decodificador.

recombinação genética é o mecanismo que reorganiza os genes já existentes no cromossomo. Nos seres humanos há possibilidade de $2^{23} = 8.388.608$ gametas distintos criados por cada indivíduo (gametogênese), em torno de 70 trilhões de zigotos por casal (fecundação). Também analisa as possibilidades de fecundação cruzada (entre indivíduos distintos da mesma espécie) e autofecundação (um indivíduo que fecunda a si mesmo).

redundância excesso. Significa repetições na teoria da informação. Nos algoritmos de compressão são os códigos disponíveis que não podem ser utilizados pois o processo de compressão ficará irreversível. A redundância do código genético é a existência de mais de um códon para a representação de um único aminoácido.

região abissal águas oceânicas a partir de 4.000 metros de profundidade. A luz solar não atinge esta região e a temperatura é constante, em torno de 2°C o ano todo. Os animais que ali habitam possuem adaptações específicas como células luminescentes.

região intergenes são segmentos contíguos do DNA que não são genes. Nos seres procarióticos, esta região tente a ser muito pequena; nos seres humanos, estima-se que ela representa 90% do DNA [36]. Também chamada de região entre os genes, lixo genético e *junk DNA*.

relações desarmônicas são as ligações nas quais há prejuízo para alguns dos seres envolvidos.

relações entre os seres vivos são as formas de interação dos seres vivos. Há cinco formas principais de relações: sociedade, mutualismo, comensalismo, predatismo e parasitismo; as três primeiras são harmônicas e as duas últimas desarmônicas.

relações harmônicas são as ligações nas quais não há prejuízo a nenhum dos seres que dela participam.

renaturação processo que reúne as cadeias polinucleotídicas do DNA imediatamente após a

síntese (no trecho) do RNA mensageiro. É o processo contrário ao da desnaturação.

replicação réplica, cópia. Auto replicação: capacidade de fazer uma cópia de si. Replicação do DNA, o mesmo que duplicação do DNA.

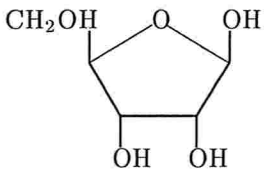
resíduos vide monômeros.

retículo endoplasmático organela que possui uma estrutura membranosa rica em ribossomos.

retinoblastoma neoplasia (tumor) ocular maligna da segunda infância. Normalmente ocorre antes do terceiro ano de vida [55].

retrovírus organismo que possui duas moléculas de RNA, fita simples. Tem a integração, como DNA fita dupla, no cromossomo das células hospedeiras como característica fundamental para sua reprodução. Além do RNA, o retrovírus possui um tipo de tRNA e várias proteínas codificadas pelo vírus. Destaca-se a *transcriptase reversa*, uma enzima que realiza a síntese da molécula de DNA fita dupla, a partir do RNA viral, necessária à integração. São hospedeiros de células eucarióticas [58].

ribose pentose do ácido ribonucléico.



ribossomos maquinaria molecular para a síntese protéica, podendo ser encontrados na membrana do retículo endoplasmático ou no interior do citoplasma. É um complexo constituído de RNA e proteínas [26].

RNA ácido ribonucléico. Cadeia polinucleotídica que possui diversos tipos cada qual com uma função distinta dentro da célula. Todos os tipos de RNA possuem alguma função no processo da síntese protéica.

RNA mensageiro (mRNA), cópia do gene que servirá como base para a síntese protéica.

RNA polimerase (RNAP), um dos responsáveis pela transcrição do mRNA. É uma enzima.

RNA ribossômico (rRNA), um dos responsáveis pela tradução do mRNA.

RNA transportador (tRNA), responsável pelo transporte dos aminoácidos para a tradução do mRNA.

sacarídeo o mesmo que carboidrato.

seleção natural é o processo no qual somente os seres mais aptos de uma espécie sobrevivem. A sorte influencia diretamente na seleção natural, que é o principal fator evolutivo atuante sobre a variabilidade genética da população. Numa explicação simples, a evolução é o resultado da atuação da seleção natural sobre a variabilidade genética de uma população. Ela age como um estabilizador, eliminando os fenótipos desviantes.

seqüência genômica vide fita.

seqüenciamento genético atividade da identificação da seqüência de nucleotídeos na cadeia codificadora de algum DNA.

seres autotróficos são os organismos capazes de produzir todas as moléculas orgânicas que necessitam, a partir de substâncias inorgânicas retiradas do ambiente. São os seres que produzem a própria energia a partir do ambiente que vivem.

seres vivos são aqueles capazes de se reproduzir, utilizar energia, consumir matérias primas, eliminar dejetos e reagir ao mundo exterior. Todo ser vivo possui seu código genético, indispensável à sua reprodução. Não existe vida sem o processamento de informação nem sem água no estado líquido (a vida pode estar latente em uma bactéria congelada, mas não estará ativa nestas condições).

significado da informação é o trato semântico da mensagem, a informação semântica.

símbolo conjunto de um ou mais caracteres que formam o alfabeto.

símbolo de fim de mensagem símbolo especial utilizado pelo algoritmo de compressão aritmética para indicar o final da mensagem.

síntese protéica processo celular que forma as proteínas necessárias para a manutenção das funções celulares.

sociedade é uma associação de indivíduos da mesma espécie que, embora possam viver isoladamente, procuram se agrupar e levar uma vida de forma coletiva. Uma colmeia, por exemplo.

sopa primordial hipotética quantidade de elementos químicos essenciais à vida existente no oceano pré-biótico, a partir da qual a vida teria originado.

superposição de genes seqüência contígua do DNA que possui a codificação para mais de uma proteína. A superposição pode ser parcial ou total.

taxa de compressão medida da compressibilidade de uma mensagem. Normalmente é quantificada em porcentagem.

tecido Conjunto de células que possuem as mesmas características funcionais [21].

técnicas de dicionários são algoritmos de compressão que utilizam um dicionário para a codificação da mensagem, que geralmente é representada pelas entradas no dicionário seguidas ou não da representação de algum símbolo.

técnicas LZ técnica de compressão cujo dicionário é composto pela memória do segmento de cadeia já codificado. O nome LZ se refere a seus criadores Jacob Ziv e Abraham Lempel. Os compressores que utilizam esta técnica são muito eficientes para a linguagem natural, como Unix Compress, Arj, etc.

teia alimentar é o entrelaçamento de diversas cadeias alimentares em um mesmo ecossistema.

teoria da evolução Defende a hipótese de que os seres vivos existentes hoje em dia provém de antepassados de aparência muito diferente da atual. Eles viveram na terra há muitos milhões de anos e foram mudando de forma por ação de diversos fatores. Alguns como o meio ambiente, as mutações genéticas e a seleção natural foram tratados neste texto. Outros como a recombinação genética, a migração e a oscilação (ou deriva) genética foram abordados superficialmente somente neste glossário. Além destes fatores há outros aspectos que tornam a teoria da evolução um assunto complexo e controverso. A mula é infértil e resulta do cruzamento de um eqüino com um jumento; se fosse fértil, nasceria uma nova espécie. Lamark propôs uma teoria sobre as variações no interior de uma espécie. Ele concluiu que espécies aparentadas tinham se desenvolvido a partir de um grupo de subespécies, uma teoria ainda aceita mas com controvérsias [20]. Nos últimos 65 milhões de anos, o meio ambiente variou muito, mas certas espécies não. A barata e a libélula figuram entre os insetos mais antigos (período Carbonífero de 360 a 286 milhões de anos atrás). Algumas baratas fósseis lembram espécies atuais [51]. Os grupos mais antigos de tubarões se extinguiram, mas alguns descendentes deles, que surgiram há 200 milhões de anos, vivem até hoje [32]. Os crocodilos pouco mudaram desde o Jurássico (de 213 a 144 milhões de anos atrás) [51]. Os crocodilos viveram antes, durante e depois dos dinossauros e ainda estão por aqui [40]. Há muitos fósseis de celacanto datados de 400 milhões de anos de idade [42]. Os animais "prontos" contrariam a teoria da evolução e da seleção natural. Por outro lado, existem espécies cujos

fósseis trazem uma cronologia de mutações graduais. Existem ainda espécies distintas que sofreram mutações genéticas na mesma época. Há períodos propícios à evolução das espécies e outros nos quais há pouca mutação genética. A grande explosão evolucionária do período Cambriano (555 a 505 milhões de anos atrás) diversificou consideravelmente as espécies de invertebrados marinhos [27]. Existe grande complexidade no processo de evolução das espécies. Não há uma teoria universalmente aceita sobre o processo evolutivo, é um assunto delicado e controverso.

teoria da informação "teoria de comunicação que trata da codificação e da transmissão de informações por meio de sinais" [41].

tetragrama conjunto de quatro letras ou palavras, conforme o contexto utilizado.

tradução processo que converte a mensagem contida na seqüências do RNA mensageiro em proteína.

transcrição é o processo de sintetização de uma molécula de RNA. É principalmente durante a transcrição que a célula exerce o controle da expressão gênica [58]. Na maioria dos casos, o principal ponto de regulação da atividade de um gene é a decisão de iniciar ou não sua transcrição, que possui três fases: início, alongamento e terminação. Ao final as moléculas de RNA são chamadas de transcritos primários. Antes de ser utilizado pela célula, cada molécula de RNA deve ser processada para se tornar um RNA maduro. Os rRNA e tRNA são processados tanto nos seres procarióticos quanto nos eucarióticos. Os mRNA são processados somente nos os seres eucarióticos. É no processamento do RNA, após a transcrição, que os íntrons são separados dos exons.

transmissão processo do transporte da mensagem codificada para o receptor.

transmissão incremental quando a transmissão ocorre bit a bit, a media que os bits são definidos pelo codificador.

transmissor aquele que faz a transmissão. Em muitos casos é também o codificador.

trigrama conjunto de três letras ou palavras, conforme o contexto utilizado.

vacúolo organela responsável pelo estoque de muitos nutrientes e outras substâncias para as células.

valor esperado é a média ponderada dos valores possíveis, se a cadeia for finita.

variância é uma medida da variação de alguma variável aleatória. Esperança matemática do quadrado dos afastamentos de uma variável aleatória em relação à média aritmética.

variável aleatória associa cada possível resultado de um experimento a um valor real.

variável aleatória bidimensional associa cada possível resultado de um experimento a dois valores reais.

vida propriedade dos seres vivos.

vírus organismos formados por uma cápsula de proteína composta por várias subunidades. Nesta cápsula há somente um tipo de ácido nucléico, nunca os dois [22].

viróides agentes infecciosos formados por uma única molécula de RNA e sem a capa protéica. Atacam principalmente as células vegetais [22].

zigoto Célula reprodutora gerada a partir da fusão de dois gametas de sexos opostos. Ovo. Normalmente são diplóides, possuindo dois jogos de cromossomos, um de cada gameta.

Referências Bibliográficas

[1] S. Adeodato, S. Tunes e A. Beccari, Existe vida lá em cima?, *Globo Ciência* **43** (1995), 49–50.

[2] G. Aguerre, Um novo nome para o planeta azul, *Super Interessante Especial* **5** (1998), 10–11, Oceanos do Mundo.

[3] G. Aguerre e C. Ângelo, Paraíso e inferno, *Super Interessante Especial* **5** (1998), 32–39, Oceanos do Mundo.

[4] J.T. Arantes, Vida no espaço, *Galileu* **106** (2000), 34–35.

[5] T.C. Bell, J.G. Cleary e I.H. Witten, *Text compression*, Prentice Hall, 1990.

[6] I. Boscov, Entrevista com Robert Ballard, *Revista Veja* **1702** (2001), 11–15, Edição de 30 de maio.

[7] D. Burnie, Aventura na ciência, *Vida*, Editora Globo, 1994, Traduzido por R. Ziegelmaier.

[8] C. R. Calladine e H. R. Drew, *Understanding DNA - the molecule and how it works*, Academic Press, 1999, Second Edition.

[9] L. Candisani e M.S.J. França, A caçada da vida no fundo do mar, *Galileu* **117** (2001), 34–37.

[10] X. Chen, S. Kwong e M. Li, A compression algorithm for DNA sequence and its applications in genome comparison, *Genome Informatics* **10** (1999), 41–50.

[11] T.H. Cormen, C.E. Leiserson e R.L. Rivest, *Introduction to algorithms*, McGraw Hill, 1990, Fifteenth Printing, 1995.

[12] M. Coutinho, A religião contra-ataca, *Galileu* **121** (2001), 29–35, Uma reportagem sobre o criacionismo.

[13] M. Crochemore e R. Vérin, Zones of low entropy in genomic sequences, *Computers and Chemistry* **23** (1999), 275–282.

[14] A.B. de H. Ferreira, *Novo dicionário Aurélio da língua portuguesa*, Editora Nova Fronteira, 1986, Segunda edição revisada e ampliada - 13ª impressão.

[15] F. Dieguez, Trapalhadas genômicas - conclusões apressadas, *Super Interessante* **163** (2001), 19.

[16] K. Drlica, *Understanding DNA and gene cloning*, John Wiley & Sons, Inc., 1997, Third Edition.

[17] W. Ebeling e C. Frömmel, Entropy and predictability of information carriers, *BioSystems 46*, Elsevier Science Ireland Ltd., 1998, pp. 47–55.

[18] F. Facchini, Origem e evolução, *O Homem*, Editora Moderna, 1997, Traduzido por R.V. Kono e S. Visconti.

[19] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner e J. Ziv, On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence, *Proc. 5th Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM-SIAM, 1994, pp. 48–57.

[20] L. Gamlin, Aventura na ciência, *Evolução*, Editora Globo, 1994, Traduzido por R. Ziegelmaier.

[21] A. Garassino, Origem e evolução, *As Plantas*, Editora Moderna, 1997, Traduzido por H. Feist.

[22] F. Gewandsznajder e S. Linhares, *Biologia hoje - volume 1*, Editora Ática, 1999, 12ª Edição.

[23] D. Grecco, Os segredos do mar, *Galileu* **102** (2000), 54.

[24] S. Grumbach e F. Tahi, Compression of DNA sequences, *Proc IEEE Symp. on Data Compression* (1993), 340–350.

[25] ———, A new challenge for compression algorithms: genetic sequences, *Inform. Processing Manage.* **30** (1994), 875–886.

[26] D. Gusfield, *Algorithms on strings, trees, and sequences computer science and computational biology*, Cambridge University Press, 1999.

[27] N. Harley e et al., Atlas visuais, *A Pré-História*, Editora Ática, 1996, Traduzido por L. Figuti.

[28] J.K. Lanctot, M. Li e E.-H. Yang, Estimanting DNA sequence entropy, *Proc. 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM-SIAM, 2000, pp. 409–418.

[29] T. Leitner, D. Escanilla, C. Franzén, M. Uhlén e J. Albert, Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis, *Proc. Natl. Acad. Sci.* **93** (1996), 10864–10869.

[30] L. Lima, Contra a discriminação da mulher, *Super Interessante* **163** (2001), 20–21, Entrevista com a geneticista Mayana Zats.

[31] D. Loewenstern e P.N. Yianilos, Significantly lower entropy estimates for natural DNA sequence, *Journal of Computational Biology* **6** (1999), no. 1, 125–142.

[32] M. Macquitty, Aventura visual, *Tubarões*, Editora Globo, 1994, Traduzido por R. Amarante.

[33] M.N. Magalhães e A.C.P. de Lima, *Noções de probabilidade e estatística*, IME-USP, 2001, terceira edição.

[34] R. Mantegna, R. Buldyrev, S. Goldberger e et al., Linguistic features of noncoding DNA sequences, *Phys. Rev. Lett.* **73** (1993), 3169–3172.

[35] S. Mateos, Nossos avós eram extraterrestres?, *Super Interessante* **168** (2001), 21, Uma breve discussão sobre a panspermia.

[36] J. Meidanis e J.C. Setubal, *Introduction do computational molecuar biology*, PWS Publising Co., Boston, 1997.

[37] P.L. Meyer, *Probabilidade - aplicações à estatística*, Livros Técnicos e Científicos Editora, 1983, Segunda Edição - Reimpressão 1984.

[38] W. Nestlehner, O outro mundo é aqui, *Super Interessante* **126** (1998), 40–49, Sobre o Habitat dos Ventos Hidrotermais.

[39] C.G. Nevill-Manning e I.H. Witten, Protein is incompressible, *Proc. Data Compression Conference*, IEEE Press, 1999, Los Alamitos, CA, pp. 257–266.

[40] D. Norman e A. Milner, Aventura visual, *Dinossauros*, Editora Globo, 1990, Traduzido por F.S. Leite.

[41] *Nova enciclopédia ilustrada Folha*, encartes semanais de março a dezembro de 1996 - Jornal Folha de São Paulo, 1996, Coletânea de Larousse, Cambridge, Oxford e Webster.

[42] S. Parker, Aventura visual, *Peixes*, Editora Globo, 1990, Traduzido por A.M. Quirino.

[43] S. Pollock, Aventura na ciência, *Ecologia*, Editora Globo, 1994, Traduzido por L.Z. Seixas.

[44] *Center for Polymer Studies*, 2000, http://polymer.bu.edu/pub/dna/.

[45] É. Rivals, J.P. Delahaye, M. Dauchet e O.Delgrange, A guaranteed compression scheme for repetitive DNA sequences, *LIFT Lille I University* (1995), technical report IT-285.

[46] C.D. Sasso, Origem e evolução, *Os Animais*, Editora Moderna, 1997, Traduzido por H. Feist.

[47] K. Sayood, *Introduction to data compression*, Morgan Kaufmann, 2000, Second Edition.

[48] T.D. Schneider, *Information theory primer*, ftp://ftp.ncifcrf.gov/pub/delila/primer.ps, 2000.

[49] C.E. Shannon, A mathematical theory of communication, *The Bell System Technical Journal* **27** (1948), 379–423,623–656.

[50] J.L. Soares, *A Terra - preservação ambiental - ar, água, solo, ecologia e asúde*, Editora Moderna Ltda, 1995, Primeiro Grau.

[51] P.D. Taylor, Aventura visual, *Fóssil*, Editora Globo, 1990, Traduzido por A.M. Quirino.

[52] J.C.A. van der Lubbe, *Information theory*, Cambridge University Press, 1997, Traduzido para o inglês por H.J. Hoeve e S. Gee.

[53] *National Center for Biotechnology Information*, 2000, http://www.ncbi.nlm.nih.gov/.

[54] *Woods Hole Oceanographic Institution*, 2001, http://www.whoi.edu/.

[55] J.B. Wyngaarden e L.H. Smith, *Cecil textbook of medicine*, Saunders, 1985, 17th edition.

[56] H.P. Yockey, *Information theory and molecular biology*, Cambridge University Press, 1992.

[57] ———, Origin of life on Earth and Shannon's theory of communication, *Computers & Chemistry 24*, Elsevier Science Ireland Ltd., 2000, pp. 105–123.

[58] A. Zaha, A. Schrank, H. B. Ferreira, I. S. Schrank, J. J. S. Rodrigues, L. P. Regner, L. M. P. Passaglia, M. L. R. Rossetti, R. M. Raupp, S. C. Silva e V. L. V. Gaiesky, *Biologia molecular básica*, Mercado Aberto, 2000, Segunda Edição - Série Ciência XXI.

[59] J. Ziv e A. Lempel, A universal algorithm for sequential data compression, *IEEE Transacitions on Information Theory, IT* **23** (1977), no. 3, 337–343.

Índice Remissivo